博士論文

Panoramic View on Genome Diversity and Evolution of Lactic Acid Bacteria

(乳酸菌ゲノムの多様性と進化に関する俯瞰的解析)

Yasuhiro Tanizawa

谷澤 靖洋

Abstract

Lactic acid bacteria (LAB) have long been associated with human culture and industrially exploited in production and preservation of food and feed for centuries. They are isolated across the world from nutrient rich environments, such as dairy products, fermented foods, plants, and animal intestines. From a taxonomic point of view, they are distributed into over 30 genera from six families under the order *Lactobacillales*. Among them, the genus *Lactobacillus* is the largest and highly heterogeneous group comprising nearly 200 species and subspecies. Recent advance of genome sequencing technologies has realized access to enormous genomic data. Particularly in the field of microbiology, genome sequences for a variety of organisms, not limited for model organisms or human pathogens, have become available, which gave rise to new opportunities for investigating diverse species. As of April 2016, NCBI Assembly Database stores more than 700 genomes for the genus *Lactobacillus*, marking the largest number except for model microorganisms and pathogenic bacteria. In particular, they include genomic data of 179 *Lactobacillus* spp. covering over 90% of its known species. The ecological characteristics of LAB and its wealth of genomic data make this microorganism particularly attractive for revealing the diversity of microbial world and their evolutionary background.

This work contains three research projects. The first two address case analyses of LAB that exhibit atypical characteristics: *L. hokkaidonensis* and the genus *Fructobacillus*. The last one addresses the development of a genome archive and annotation pipeline specialized for LAB.

Psychrotolerant LAB: *Lactobacillus hokkaidonensis*

Lactobacillus hokkaidonensis is an obligate heterofermentative LAB, which was isolated from Timothy grass silage in Hokkaido, a subarctic region of Japan. This bacterium is considered useful as a silage starter culture in cold regions because of its remarkable psychrotolerance; it can grow at temperatures as low as 4 °C. To elucidate its genetic background, particularly in relation to the source of psychrotolerance, I reconstructed the complete genome sequence of *L. hokkaidonensis* LOOC260^T using the PacBio single-molecule real-time sequencing technology.

The genome of LOOC260^T comprises one circular chromosome (2.28 Mbp) and two circular plasmids: pLOOC260-1 (81.6 kbp) and pLOOC260-2 (41.0 kbp). I identified diverse mobile genetic elements, such as prophages, integrative and conjugative elements, and conjugative plasmids, which may reflect adaptation to plant-associated niches. I also identified unique regions of the genome and found several factors that may contribute to the ability of *L. hokkaidonensis* to grow at cold temperatures.

Fructophilic LAB: *Fructobacillus*

Fructobacillus spp. belong to the family *Leuconostocaceae* and are frequently found in fructose-rich niches, such as flowers, fruits, and bee guts. They were originally classified as *Leuconostoc* spp., but were later grouped into a novel genus, *Fructobacillus*, based on their phylogenetic position, morphology and specific biochemical characteristics. The fructophilic characteristic, referring to its preference for fructose over glucose under anaerobic conditions, has not been reported in other groups of LAB, suggesting the unique evolution at the genome level. I conducted comparative analysis using five draft genome sequences of *Fructobacillus* spp. and *Leuconostoc* spp. to reveal their adaptive evolution to the fructose-rich environments.

Compared to *Leuconostoc* spp., *Fructobacillus* species have significantly smaller number of protein coding sequences in their smaller genomes, especially lacking genes for carbohydrate transport and metabolism. Asymmetric distribution of conserved genes in each genus also shows that *Fructobacillus* spp. have lost more genes rather than have acquired new genes, indicating the streamlined genomes of *Fructobacillus* spp. The lack of *adhE* genes in all *Fructobacillus* spp. exemplified the relevance of this gene in fructophilic characteristic, as postulated in previous studies. I revealed the general trend of reductive evolution, especially in metabolic simplification based on sugar availability, in this species.

LAB genome archive and annotation pipeline

The number of LAB genomes available is drastically increasing, together with the spectrum of data quality and taxonomically mislabeled entries. They may lead to incorrect assumption and erroneous conclusions when dealt without careful consideration. In particular, some LAB species are difficult to distinguish only by the 16S rRNA gene-based identification, and a significant number of LAB genomes were deposited with incorrect taxonomic names in public databases. To resolve these issues, I developed a curated genome repository DAGA (DFAST Archive of Genome Annotation) to provide reliable genome data resources for LAB.

DAGA currently provides 1,421 LAB genomes covering 191 species/subspecies of two genera *Lactobacillus* and *Pediococcus* in the family *Lactobacillaceae* obtained from both DDBJ/ENA/GenBank and Sequence Read Archive. All genomes deposited in DAGA were re-annotated consistently using the identical pipeline. I used the average nucleotide identity (ANI), which showed high discriminative power to determine whether two genomes belong to the same species, to confirm the taxonomic position. As a result, 155 mislabeled or unidentified genomes were assigned their correct taxonomic names and 38

genomes were marked as ‘poor quality’. In particular, genomes for six type strains were disqualified due to possible misidentification or contamination. DAGA will improve both accessibility and reusability of genomic data for LAB.

To provide consistent annotation to genomes stored in DAGA, I developed an annotation pipeline called DFAST (DDBJ Fast Annotation and Submission Tool, https://dfast.nig.ac.jp), with curated reference protein databases tailored for LAB as well as quality and taxonomy assessment methods. DFAST was developed so that all the procedures required for data submission could be performed seamlessly online, and it can generate ‘ready-to-submit’ level annotation files to DDBJ without computational knowledge.

By exploiting the data deposited in DAGA, I found previously unreported intraspecific diversity within *Lactobacillus gasseri* and *Lactobacillus jensenii* that might deserve subspecies-level differentiation. In addition, through the analysis of gene transfer among LAB strains, the niche-specific dissemination of genes related to anti-stress system was identified.

Contents

Journal publication

This dissertation includes the contents of three publications with peer review process listed below.

1. **Tanizawa, Y.**, **Tohno, M.**, **Kaminuma, E.**, **Nakamura, Y.**, **Arita, M.** Complete genome sequence and analysis of *Lactobacillus hokkaidonensis* LOOC260^{T}, a psychrotrophic lactic acid bacterium isolated from silage. *BMC Genomics* **16,** 240 (2015).
2. **Endo, A.**$^{\#}$, **Tanizawa, Y.**$^{\#}$, **Tanaka, N.**, **Maeno, S.**, **Kumar, H.**, **Shiwa, Y.**, **Okada, S.**, **Yoshikawa, H.**, **Dicks, L.**, **Nakagawa, J.**, **Arita, M.** Comparative genomics of *Fructobacillus* spp. and *Leuconostoc* spp. reveals niche-specific evolution of *Fructobacillus* spp. *BMC Genomics* **16,** 1117 (2015). # Contributed equally.
3. **Tanizawa, Y.**, **Fujisawa, T., Kaminuma, E.**, **Nakamura, Y.**, **Arita, M.** DFAST and DAGA: Web-based integrated genome annotation tools and resources. *Bioscience of Microbiota, Food and Health* **35** (2016, in press).

Additionally, I published five articles without peer review process during my doctoral program.

1. **Tanizawa, Y.**, **Fujisawa, T.**, **Mochizuki, T.**, **Kaminuma, E.**, **Suzuki, Y.**, **Nakamura, Y.**, **Tohno, M.** Draft genome sequence of *Weissella oryzae* SG25^{T}, isolated from fermented rice grains. *Genome Announcements* **2,** e00667–14 (2014).
2. **Tanizawa, Y.**, **Fujisawa, T.**, **Mochizuki, T.**, **Kaminuma, E.**, **Nakamura, Y.**, **Tohno, M.** Draft genome sequence of *Lactobacillus oryzae* strain SG293^{T}. *Genome Announcements* **2,** e00861–14 (2014).
3. **谷澤靖洋** ANI は菌種同定に使えるか? *日本乳酸菌学会誌* **26,** 206 (2015).
4. **谷澤靖洋**, **神沼英里**, **中村保一**, **清水謙多郎**, **門田幸二** 次世代シーケンサーの解析手法第６回ゲノムアセンブリ *日本乳酸菌学会誌* **27,** 41–52 (2016).
5. **谷澤靖洋**, **神沼英里**, **中村保一**, **遠野雅徳**, **大崎研**, **清水謙多郎**, **門田幸二** 次世代シーケンサーの解析手法第７回ロングリードアセンブリ *日本乳酸菌学会誌* **27,** 101–110 (2016).

Chapter 1

1. General Introduction

1.1. Advance of genomic studies in microbiology

1.1.1. *The era of Genome sequencing*

The term 'genome' was originally coined by German botanist Hank Winkler in 1920 as the combination of the word 'gene' and the suffix '–ome' that refers to a totality of some sort [1]. The word 'gene' had already introduced by Danish biologist Wilhelm Johannsen in 1909, and DNA had also been discovered as early as in the late 19th century. However, the relationship between gene and DNA was not established at that time. It was only after a series of experiments starting from the late 1920s conducted by Griffith, Avery and his colleagues, and Hershey and Chase that DNA was accepted as the genetic material in the 1950s [2]. Now 'genome' is generally used as a word denoting the full complement of DNA in an organism. The word genome also implies completeness because all genetic information contained in a single genome mostly determines life patterns of its organism [3]. The year 1953 marked the dawn of molecular biology by the discovery of the double helical structure of DNA by James Watson and Francis Crick [4]. Ever since the advancement in molecular biology had laid the foundation to develop the method for determining the sequence of nucleotides in a DNA molecule.

The modern sequencing technology was developed in 1970s through the two independent studies: the Maxam-Gilbert method [5] and the Sanger method [6,7]. The Sanger method, based on the chain-termination using dideoxynucleotides (ddNTPs), became popular owing to its relative convenience against the Maxam-Gilbert chemical cleavage method. Later, the Sanger method was improved and automated by the incorporation with fluorescent dye-labeled ddNTPs and capillary electrophoresis, and dominated the DNA sequencing during the subsequent 30 years [8]. Automated Sanger sequencing was also employed in the Human Genome Project and, together with the whole-genome shotgun sequencing method, gave a boost to the completion of the project [8-11].

As of writing this thesis, twenty years have already passed since the first complete genome for a cellular organism, *Haemophilus influenzae,* was reported in 1995 [12]. The first and latter decades differ dramatically. In the first decade, the sequenced genomes were mostly limited to those for model organisms like *Escherichia coli*, *Bacillus subtilis*, and *Saccharomyces cerevisiae*, or those of special

importance medically, industrially, or economically with considerable biases to human pathogens [8,13-15]. The large sequencing centers or international consortia took the initiative in many of the sequencing projects in the early stage of the decade, and automated Sanger sequencers were predominantly used [16].

When the massively parallel sequencing technology was introduced in 2005, the situation was dramatically changed. It is also referred to as 'second- or next- generation sequencing' (NGS) in contradistinction to the capillary-based Sanger sequencing method which is considered as the 'first-generation'. NGS can generate enormous volume of data excessing million to billion reads per sequencing run and drastically reduced the cost and labor for DNA sequencing, thereby enabling single research teams to conduct their own sequencing projects routinely [17,18]. Currently, the number of bacterial genome projects deposited in the Genomes OnLine Database (GOLD) exceeds 50,000, which increased 50-fold over 10 years since 2005 [19]. Taking the advantage of NGS technologies, large scale sequencing projects were launched. For instance, the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project aims to collect comprehensive genome sequences mainly for type strains from all cultured bacteria and archaea [20].

More recently, third-generation sequencing technology has emerged. Even with NGS, finishing a bacterial genome remains a costly and time-consuming process. Therefore many of the genomes produced by NGS have been published in draft status. The third-generation sequencers, such as PacBio and Nanopore [21,22], featuring much longer reads from single-molecule DNA compared with NGS, can facilitate assembly of complex and repetitive regions of the genome and thus have potential to generate finished genomes at much lower cost [23].

Currently, the third-generation has not fully replaced the second-generation. Although NGS no more stands for 'next-generation', the impact of NGS upon biology is revolutionary and the term 'NGS' is still prevalently used as the representative of state-of-art sequencing technology. The trend that both generations' technologies are used in parallel will continue, as they exhibit distinctive characteristic, high-throughput and long-read, respectively. In any case, now is the generation that access to enormous genomic data has become realized, which by no means could be accomplished in the past generation.

1.1.2. *Trends in bacterial genome analysis*

The increased availability of whole genome sequences provided new opportunity for microbiology. The pan-genomic study is one of such studies made feasible by the low sequencing cost that rendered sequencing hundreds of strains in a species or a genus affordable. The pan-genome is defined as the full genetic repertoire in a given clade, which is comprised of a core genome shared by all strains in that clade

and an accessory (or dispensable) genome shared by some but not all strains [24,25]. The pan-genome can be either 'open' or 'closed'. The first pan-genomic study for *Streptococcus agalactiae* and the one using 186 *E. coli* genomes showed that the core genome would constitute only a small fraction of the pan-genome while the pan-genomes for these species were 'open', i.e. the size of the pan-genome will continue to increase when adding a new genome, indicating the vast and diverse gene pool behind them. These studies emphasize the importance of sequencing multiple strains to capture the diversity of bacterial species [24,26].

Identification of the core genome also facilitates the robust method for phylogenetic reconstruction: the core genome phylogeny or 'phylogenomics' approach based on the comparison of the whole shared genome information. Since late 1980s, when DNA sequencing became commoditized, sequence-based molecular phylogenetics has advanced, among which the method using 16S rRNA gene sequence as a molecular marker has established the standard criteria in prokaryotic systematics [27]. However, such single-gene-based methods often yield inconsistent results depending on the genes analyzed, or phylogenetic signal obtained from single gene is sometimes insufficient to distinguish particular species. In general, phylogenetic trees constructed from multiple genes can exhibit finer resolution and more robustness against horizontal gene transfer in resolving phylogenetic relationships below the phylum level and provide an excellent schema for bacterial phylogeny [28,29].

Apart from these single/multiple-gene based methods, several new attempts have been proposed to measure genetic relatedness based on whole-genome sequences [30-34]. In particular, average nucleotide identity (ANI) is most widely acknowledged due to its simplicity and robustness. ANI is calculated from the mean sequence identity of homologous regions in the pair-wise comparison of two genomes. In the current systematics, DNA-DNA hybridization (DDH) value of 70% still remains as 'gold standard' for describing a bacterial species, albeit this classical demarcation was proposed almost 30 years ago [35]. ANI correlates well with DDH as well as 16S rRNA gene similarity, where ANI values of 95-96% correspond to DDH values of 70%; when the ANI value between two genomes shows higher than this threshold, the two genomes can be considered to belong to the same species [31,32]. ANI has already been used to describe new species as a replacement or a complement of tedious and complicated DDH method [36-38]. Recently, several new algorithms for improved ANI calculation were proposed, although their significance has yet to be determined [39,40]. It was argued that the combination of core-genome-based phylogeny and ANI provides an appropriate method for bacterial species delineation [41].

Another strategy to reveal the microbial diversity is culture-independent sequencing of DNA directly extracted from environmental samples, which is commonly referred to as metagenomics. Metagenomics

has shed light on unculturable strains in microbial communities, which are estimated to dominate over 99% of the all microorganisms in natural environments. In contrast to the limited number of metagenomic studies described in the first sequencing generation [42-44], the advent of NGS made metagenomics more practicable with the ability to generate high-throughput data to capture low-abundance organisms present in the sample. 16S rRNA amplicon sequencing and shotgun metagenomic sequencing are both effective approaches to characterize the diversity of microbial populations. The former mainly focuses on the composition of the communities, while the latter focuses on functional aspects as well [45]. Although challenging, functional metagenomics would be a prospective method that can identify novel genes and pathways beneficial for biotechnological use or can reveal the interaction and co-evolution between host and microbiome [46]. During the last decade, large scale metagenome projects have been carried out, for example, the TerraGenome project for soil, and the MetaHit project and the Human Microbiome Project for human intestinal microbiome [16]. In recent studies, an expanded view of the tree of life was proposed by exploiting over 1,000 genome data reconstructed from environmental samples, where the vast diverse groups of uncultivated and little known organisms dominated a large part of the bacterial domain [47,48].

1.1.3. *Public sequence database*

In 1960s, the recognition of importance of molecular sequences in biological contexts gave rise to the attempt to collect and archive sequence data. Margaret Dayhoff pioneered such effort and published the series of Atlas of Protein Sequence and Structure from 1965 to 1978, which became a predecessor of the Protein Information Resource, the oldest protein sequence database established in 1984 [49].

In 1979 at the Rockefeller University in New York, a conference was held to discuss the necessity to create a centralized nucleotide sequence database. Meanwhile in Europe, the European Molecular Biology Laboratory (EMBL) held its own workshop to discuss the establishment of a sequence database, and one year later from the workshop, the EMBL Data Library (now European Nucleotide Archive, ENA) was founded in 1981 as the first central depository of nucleotide sequence data in the world. In 1981, almost three years later from the Rockefeller meeting, the National Institute of Health (NIH) issued a request for proposals to develop a comprehensive sequence database. The two groups, a National Biomedical Research Foundation team led by Dayhoff and a Los Alamos Scientific Laboratory team led by Walter Goad, competed for the contract, and finally Goad was awarded the contract with NIH and established the database at Los Alamos in 1982, which was later called GenBank and transitioned to the National Center for Biotechnology Information (NCBI) [50]. In his proposal, Goad claimed that the submission of sequences to the database should be mandatory for publication of an article in the scientific

journal, which well suited the scientific reward system and made large-scale data collection successful, as publication was the main incentive and reward to many experimental scientists [51]. He also proposed free access and free distribution of the data, and asserted no proprietary rights. Goad's idea that published knowledge should belong to the community as a whole later became the basis of open access to scientific knowledge [50].

Soon after the establishment, GenBank started collaboration with EMBL to exchange and share each data, and later the DNA Data Bank of Japan (DDBJ) joined the collaboration in 1986. The collaboration between the three databases is now called International Nucleotide Sequence Database Collaboration (INSDC). The INSDC developed standard format and protocol for data sharing, which enables the sequence data submitted to one of the database to be exchanged between others on a daily basis [52].

The advancement of sequence technology brought about the necessity for new categories in the sequence database: the Trace Archive for raw sequencing data mainly from gel/capillary sequencers and the Sequence Read Archive (SRA) for raw sequence reads from NGS [53]. As of March 2016, the number of bases deposited in SRA exceeds over 3,000-fold higher than that deposited in the core INSDC databases (DDBJ/EMBL-Bank/GenBank). The BioProject database and the BioSample database were also developed to describe more detailed metadata accompanying sequence data and to effectively link the information about research projects, biological source materials and sequences [54]. These databases hosted by INSDC partners constitute the foundation for accessibility, reproducibility, and reusability of scientific data. They serve both as an archival database of sequence data for the scientific literature and as a reference database for the research community [55].

Apart from the 'primary databases' described above, 'secondary databases' were also constructed for specific purposes. The Reference Sequence (RefSeq) is a collection of non-redundant and curated sequences from INSDC archives operated by NCBI. The RefSeq collection for prokaryotic genomes provides more than 40,000 genomes from a wide range of organisms as of July 2015 [56]. Formerly, only selected genomes of representative strains were accepted into RefSeq. However, the scope of RefSeq has changed to include all prokaryotic genomes submitted to INSDC that passed minimum quality control in order to provide consistent annotation to many draft genomes submitted without annotation [57]. The NCBI Assembly database is another resource for stable accessioning and data tracking for genome assembly data. It bundles multiple entries comprising each genome assembly, such as sequences for a chromosome and plasmids in a complete genome or multiple contigs/scaffolds in a draft genome, and then issues unique identifiers with version numbers for them, which facilitates an easy access to a specific version of a genome assembly [58].

According to the INSDC policy, the assurance of data quality and accuracy of the description are the submitter's responsibility. The explosion of the data amount resulted in the spectrum of data quality in the sequences deposited in the databases. Currently, 90% of the bacterial genome sequences newly submitted to INSDC are draft genomes. In many cases, even draft genomes are sufficient for comparative analysis based on the gene content [16]. However, it was also pointed out that 10% of the draft genomes showed poor quality to use as exemplified by the quality assessment of 32,000 publicly available bacterial genomes obtained from public databases [59]. Furthermore, taxonomically mislabeled entries have become serious concern, as they may lead to incorrect assumption and erroneous conclusion when dealt without careful consideration [55]. These mislabeled entries may result from either incorrect identification of the strain, sample contamination, sample mix-ups, or subsequent taxonomic reclassification after data submission. To address this issue, taxonomic positions of genomes deposited in RefSeq are validated by using a 16S rRNA gene or ribosomal protein genes [60]. The EzTaxon database provides curated taxonomic information based on the comparison of 16S rRNA. Recently, the use of genomic comparison methods such as ANI was also proposed for taxonomic validation [55,61,62].

1.2. Lactic acid bacteria

1.2.1. *Application and taxonomy*

Lactic acid bacteria (LAB) have long been associated with human culture and industry. They are functionally defined as the group of Gram-positive, anaerobic or microaerophilic, non-sporeforming, low GC content microorganisms that produce mainly lactic acid as end-products of carbohydrate fermentation. They are nutritionally fastidious organisms which prefer carbohydrate- and protein-rich environments, such as milk, meet, vegetable as well as animal intestinal tracts or oral cavities [63,64] . Their industrial application owes much to the metabolic property to produce lactic acid; the acidic and anaerobic conditions prevent the proliferation of food spoilage microorganisms, making LAB dominant in food microflora. They have been exploited in production and preservation of food and feed for centuries, and more recently, their application as probiotics has been attracting more and more attention for their role in health promotion or immunomodulation in the gastrointestinal tract [65-67]. Due to their long history of safe use in food and food production, many of the LAB are designated as GRAS (generally recognized as safe) by the American Food and Drug Administration (FDA). It is estimated that the market value of dairy food products associated with LAB including probiotic products reaches more than 100 billion Euros [68]. LAB are also exploited in the production of various traditional fermented foods and beverages

especially in Asia, in which LAB activities not only contribute to preservation but confer characteristic flavor and context. In livestock industry, LAB promote silage fermentation, a natural biopreservation process for forage, and/or improve digestibility by cattle. Besides, the ability to produce various metabolites makes LAB promising candidates for biotechnological use both in food and non-food industry: antimicrobial molecules like bacteriocins, food complements like vitamin or γ-aminobutyric acid (GABA), biorefineries from plant-derived biomass, and biodegradable plastics [65,69].

From a taxonomic point of view, LAB belong to the order *Lactobacillales* under the class *Bacilli* of the phylum *Firmicutes*. In the beginning of 20th century, LAB were classified into four genera: *Lactobacillus*, *Pediococcus*, *Leuconostoc*, and *Streptococcus*. Now they are expanded to about 30 genera distributed in six families: *Aerococcaceae*, *Carnobacteriaceae*, *Enterococcaceae*, *Lactobacillaceae*, *Leuconostoccaceae* and *Streptococcaceae* [70,71]. Figure 1.1 shows the phylogenetic tree of 26 representative LAB species from 12 genera in 6 families. Of note, although the genus *Bifidobacterium* is often considered as a member of LAB due to its common habitat and health-promoting role in intestinal tracts together with LAB, it is placed in a different taxon from genuine LAB at the phylum level [72]. Among them, the family *Lactobacillaceae* is the largest and highly heterogeneous group, which contains the genus *Lactobacillus* comprising more than 180 species and subspecies. According to the Bergey's Manual of Systematic Bacteriology, the genus *Lactobacillus* is further classified into three subgroups based on the fermentation properties: group I for obligate homofermentative species, group II for facultative heterofermentative species, and group III for obligate heterofermentative species [73]. Recent updated phylogenetic analysis divided them into 15 or more groups based on 16S rRNA gene sequences, although no valid subgenus-level classification is established [63]. The genus *Pediococcus* is another member of *Lactobacillaceae* with a characteristic feature of being a tetrad-forming coccus. However, it is phylogenetically placed within the *Lactobacillus* cluster, suggesting that it may be considered as part of the genus *Lactobacillus* [74]. The term *Lactobacillus sensu lato* is also proposed to refer to both genera [75]. The number of new species described for both two genera has been growing in recent years with the improvement of isolation, cultivation, and identification methods (solid line in Figure 1.2).

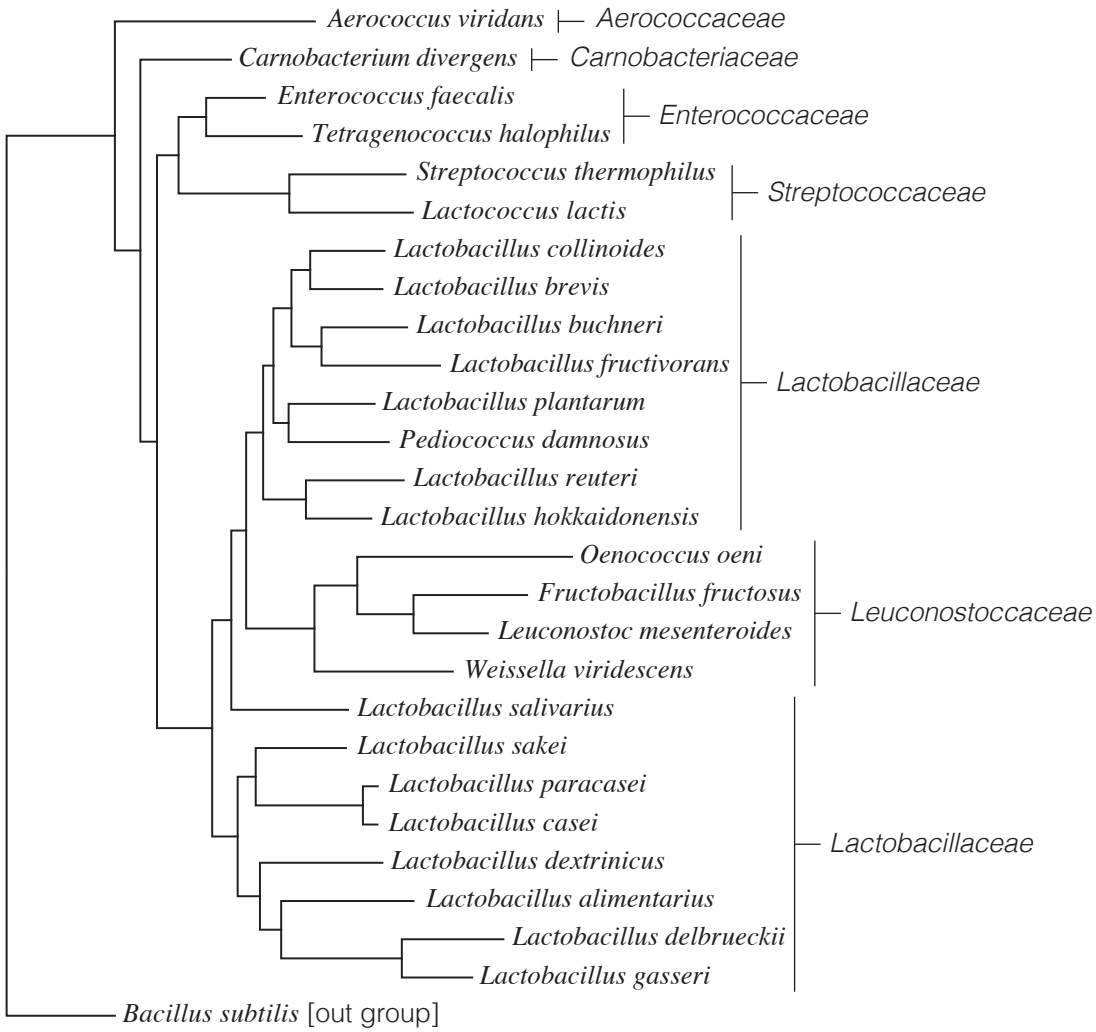


Figure 1.1. Phylogenetic tree of 26 representative LAB species from 12 genera in 6 families.

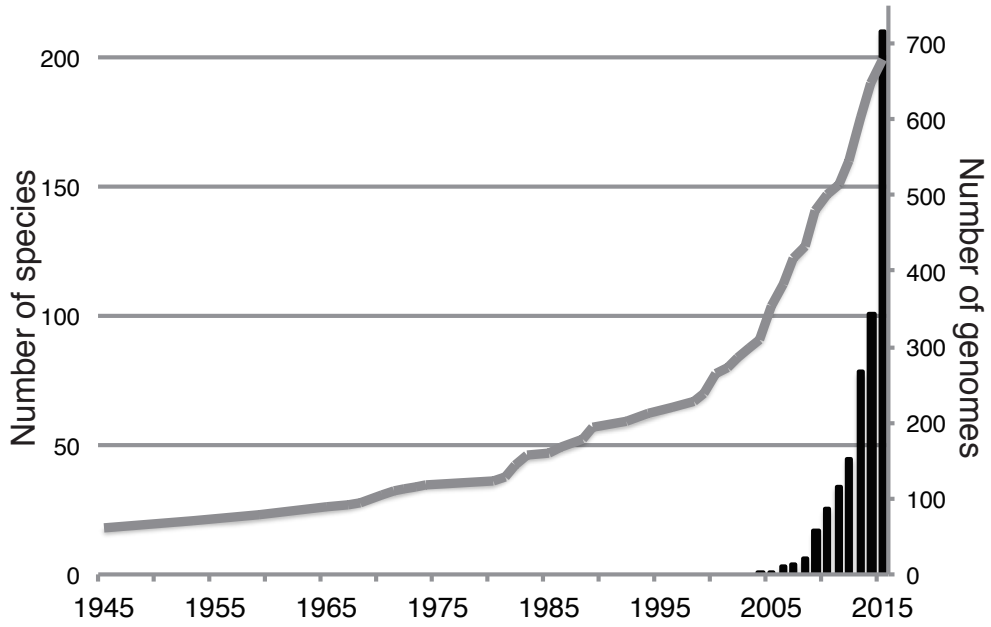


Figure 1.2. The number of described species and published genomes in *Lactobacillus* and *Pediococcus*. Solid line represents the cumulative number of described (sub-) species. Only valid species as of January 2016 were included, not reclassified ones. The bar chart represents the cumulative number of genomes deposited in DDBJ/ENA/GenBank.

The modern molecular phylogenetic method also led to many taxonomic revisions. Such representative examples include *Pediococcus dextrinicus*, which was originally identified as *Pediococcus* due to its spherical cell morphology. It was later reclassified into the genus *Lactobacillus* based on the 16S rRNA gene similarity and marks an atypical 'coccoid' lactobacillus [76]. Many of the new genera have also been described since the 1990s. For example, *Tetragenococcus* and *Oenococcus* were each assigned as a distinct genus transferred from *Pediococcus* and *Leuconostoc*, respectively [77,78]. *Fructobacillus* is a newly described genus reclassified from *Leuconostoc* in 2008, which is also known as a atypical 'fructophilic' LAB due to its preference for fructose over glucose under anaerobic conditions [79].

On the other hand, 16S rRNA gene-based method does not exhibit enough resolution to distinguish species in certain groups, such as *L. casei* group and *L. plantarum* group, in which nucleotide similarities exceed 99% with their close relatives [63]. In particular, there had been a matter of extensive debate over the taxonomy of *L. casei-paracasei* complex for more than 10 years. As a consequence of a long controversy including the redesignation of the neotype strain, majority of the strains denoted as *L. casei* today are in

fact members of *L. paracasei*. In spite of its renowned name with long history and wide-spread use in the production of fermented dairy beverage, the name *L. casei* is restricted for only a few strains [80,81].

1.2.2. *Genomics of lactic acid bacteria*

The genome sequence era of lactic acid bacteria began in the early 2000s with the complete genome sequence of *Lactococcus lactis* subsp. *lactis* IL1403 in 2001 followed by *Lactobacillus plantarum* WCFS1 in 2003 [82-84]. Both genomes are now acknowledged as reference genomes of LAB. The number of genomes available in public databases has been rapidly increasing from that time on (bar chart in Figure 1.2).

First comparative studies were reported in 2006 independently by two groups. Canchaya et al. conducted the phylogenomic analysis based on the whole genome information of five strains of *Lactobacillus* [85], and Makarova et al. revealed the regressive evolution of LAB with extensive gene loss during the diversification from their common ancestor using 12 strains from six genera [64]. Later in 2010, Kant et al defined orthologous genes shared by 20 complete genomes of *Lactobacillus* and discussed their distribution among subgeneric groups [86]. The emergence of high-throughput parallel sequencing technology in the late 2000s produced hundreds of genomes at unprecedented speed and at much lower cost, bringing about the paradigm shift in genome analyses. The comparative analysis using 100 *Lactobacillus rhamnosus* strains and pan-genomic study using 191 *Oenococcus oeni* strains represent good instances in the new generation sequencing era [87,88]. Nowadays, most type strains have been sequenced and become publicly available through large-scale sequencing projects, such as "Genome sequencing of JCM strains under the NBRP program" in Japan (BioProject ID: PRJDB547), "*Lactobacillus* in severe early childhood caries" by the Sanger Institute in UK (PRJEB3060), and "Genomic characterization of the genus *Lactobacillus*" in China (PRJNA222257). The results of such projects realized comprehensive analyses covering almost 90% of the known species based on the genomic information. Zheng et al. conducted phylogenomic and metabolic analysis using 174 type strains of *Lactobacillus* and *Pediococcus* [75]. Sun et al. used 213 strains from *Lactobacillus* and related genera for mining genes important for biotechnological application such as genes involved in carbohydrates and protein modification, cell surface interaction, and CRISPR-Cas system [89].

1.3. Organization and purpose of the dissertation

As described previously, enormous genomic data for LAB have become available today thanks to the

advance of genome sequencing technologies and large scale sequencing projects. Notably, as of April 2016, NCBI Assembly Database stores more than 700 genomes for the genus *Lactobacillus*, marking the largest number except for model microorganisms and pathogenic bacteria. The ecological characteristics of being isolated from nutrient rich environments all across the world including dairy products, fermented foods, plants, and animal intestines; wide-spread use in food and health industry; as well as the wealth of genomic data covering over 90% of its known species make this microorganism attractive research subject. I employed LAB as a genome model to unveil the diversity of the microbial world and their evolutionary background.

This dissertation contains three major chapters. The first two address case analyses of LAB that both exhibit atypical characteristics: *L. hokkaidonensis* and the genus *Fructobacillus*. The last one addresses the development of a genome archive and annotation pipeline specialized for LAB, aiming to establish an integrated research platform that makes accurate and more rapid genome analysis and to deal with more and more genomic data expected to emerge in the near future.

In chapter 2, genome analysis of a psychrotolerant LAB, *Lactobacillus hokkaidonensis*, is described. *L. hokkaidonensis* is an obligate heterofermentative LAB, which was isolated from Timothy grass silage in Hokkaido, a subarctic region of Japan. This bacterium is considered useful as a silage starter culture in cold regions because of its remarkable psychrotolerance; it can grow at temperatures as low as 4 °C [90]. To elucidate its genetic background, particularly in relation to the source of psychrotolerance, I reconstructed the complete genome sequence of *L. hokkaidonensis* LOOC260^{T} using the PacBio single-molecule real-time sequencing technology. The whole genome sequence obtained using long reads from third-generation sequencing technology enabled a genome-wide perspective of mobile genetic elements such as plasmids, prophages, and integrated and conjugative elements, which may reflect adaptation to plant-associated niches. I also identified unique regions of the genome and found several contributing factors to the ability of *L. hokkaidonensis* to grow at cold temperatures.

In chapter 3, comparative analysis of the genera *Fructobacillus* and *Leuconostoc* is described. *Fructobacillus* spp. belong to the family *Leuconostocaceae* and are frequently found in fructose-rich environments, such as flowers, fruits, or bee guts. They were originally classified as *Leuconostoc* spp., but were later grouped into a novel genus, *Fructobacillus*, based on their phylogenetic position, morphology and specific biochemical characteristics [79]. The unique fructophilic characteristic, referring to its preference for fructose over glucose under anaerobic conditions suggests its unique evolution at the genome level. I employed five draft genome sequences of *Fructobacillus* spp. for comparison with *Leuconostoc* spp. in order to reveal their adaptive evolution in the fructose-rich environments. The

analysis of conserved genes in each genus and comparative functional genomics clearly indicated the reductive evolution of *Fructobacillus*, especially in metabolic simplification based on sugar availability.

In chapter 4, the development of the genome archive and annotation pipeline specialized for LAB is described. The increasing number of genomes available in public databases resulted in the spectrum of data quality and taxonomically mislabeled entries, which may lead to incorrect assumption and erroneous conclusions when dealt without careful consideration. In particular, some LAB species are difficult to distinguish only by the 16S rRNA gene-based identification, and a significant number of LAB genomes were deposited with incorrect taxonomic names in public databases [81,91].

To resolve these issues, I developed a curated genome repository DAGA (DFAST Archive of Genome Annotation, https://dfast.nig.ac.jp) to provide reliable genome data resources for LAB. DAGA collected genomic data from both DDBJ/ENA/GenBank and SRA and their taxonomic affiliation and data quality and were assessed by using ANI and inspecting the presence of specific gene markers by CheckM [92], respectively. All genomes deposited in DAGA were re-annotated consistently using the identical pipeline called DFAST (DDBJ Fast Annotation and Submission Tool) with curated reference protein databases tailored for LAB. DFAST was developed so that all the procedure required for data submission can be performed seamlessly online, and it can generate annotation files to DDBJ without computational knowledge. By exploiting the data deposited in DAGA, exploration of intraspecific diversity within LAB and the gene transfer among LAB strains are also described.

Chapter 2

2. Genome analysis of *Lactobacillus hokkaidonensis*

2.1. Introduction

Silage fermentation is promoted mainly by the microbial activities of lactic acid bacteria (LAB). During the fermentation process, LAB produce lactic acid anaerobically as the major end product of central carbohydrate metabolism, which reduces the pH of the surrounding environment. These anaerobic and acidic conditions prevent the propagation of detrimental microorganisms such as listeria, clostridia, yeasts, and other fungi. However, the acid production level tends to be insufficient if silage is prepared in cold weather conditions because of the impaired activity of LAB, thereby yielding lower quality silage. Therefore, the inoculation of appropriate LAB as a silage additive is required to enhance silage fermentation in low-temperature environments.

Lactobacillus hokkaidonensis was a novel psychrotrophic *Lactobacillus* species isolated from Timothy grass (*Phleum pratense*) silage in Hokkaido, a subarctic region of Japan [90]. *L. hokkaidonensis* can grow at temperatures as low as 4°C (optimal growth at 25°C), and its type strain LOOC260^T was shown to decrease pH even in cold conditions when used to inoculate pilot-scale grass silage. Thus, *L. hokkaidonensis* is expected to be suitable for use as an effective silage inoculant in cold regions.

L. hokkaidonensis is classified as an obligate heterofermentative LAB in the *L. vaccinostercus* group [93], which includes five species (*L. vaccinostercus* [94], *L. suebicus* [95], *L. oligofermentans* [96], *L. nenjiangensis* [97], and *L. hokkaidonensis*) that form a clade distinct from the well-known heterofermentative clades, which include *L. reuteri*, *L. brevis*, and *L. buchneri*. They share common phenotypic features such as the presence of meso-diaminopimelic acid in their peptidoglycan cell walls and faster assimilation of pentoses compared with hexoses, but little is known about their genetic background or genomic information.

In the present study, whole-genome sequencing of *L. hokkaidonensis* LOOC260^T and comparative genome analysis were performed with emphasis on the unique gene repertoire of the *L. vaccinostercus* group. In addition, determining the complete genome may provide a better genome-wide understanding of mobile genetic elements, thereby highlighting how flexible genome rearrangements contribute to adaptation to various ecological niches. The aim of this study is to gain insights into the genomic features

of the *L. vaccinostercus* group, which is poorly characterized at present, as well as to clarify the silage fermentation mechanism from a genomic perspective, particularly in cold conditions.

2.2. Methods

2.2.1. *Genome sequencing and de novo assembly*

The cells of *L. hokkaidonensis* LOOC260^T were cultured in MRS (de Man, Rogosa, and Sharpe) broth (Difco) and were harvested in the mid-logarithmic phase. The genomic DNA was extracted and purified using Qiagen Genomic-tip 500/G and Qiagen Genomic DNA Buffer Set with lysozyme (Sigma) and proteinase K (Qiagen) according to the manufacturer's instruction. PacBio SMRT whole-genome sequencing was performed using a PacBio RSII sequencer with P4-C2 chemistry. Four SMRT cells were used for sequencing, thereby yielding 163,376 adapter-trimmed reads (subreads) with an average read length of approximately 4 kbp, which corresponded to approximately 250-fold coverage. *De novo* assembly was conducted using the Hierarchical Genome Assembly Process (HGAP) software implemented in the SMRT Analysis package 2.0, which yielded seven contigs. Independent genome sequencing using the 250-bp paired-end Illumina MiSeq system generated 5,942,620 reads, which were assembled into contigs using Platanus assembler ver 1.2 with the default settings [98]. The initial contigs derived from the HGAP method were inspected to determine their continuity with each other based on comparisons with the contigs obtained from the Platanus assembler, and were concatenated into one closed circular chromosome and two circular plasmids. The genome obtained was mapped with reads obtained by the MiSeq system using Burrows-Wheeler Alignment tool (BWA) ver 0.7.5 to detect any assembly and sequence errors [99]. As a result, six one-base-length indels were corrected. The replication origin of the chromosome (*oriC*) was predicted using the Automated Prediction Of Bacterial Replication Origin (APBRO) tool [100], and the chromosome was adjusted so the first base was upstream of the *dnaA* gene in the *oriC* region.

2.2.2. *Plasmid copy number estimation*

The plasmid copy numbers were calculated based on the read depth mapped onto each replicon. The reads obtained by the MiSeq system were mapped onto the assembled genome sequences using BWA, and the number of reads mapped onto each replicon was normalized by dividing by its sequence length. The plasmid copy numbers were determined based on the ratio of normalized read numbers for the

plasmids relative to that for the chromosome.

2.2.3. *Genome annotation*

The genome was annotated using the Microbial Genome Annotation Pipeline (MiGAP) [101] and some of the results were manually curated. In the pipeline, protein coding sequences (CDSs) were predicted by MetaGeneAnnotator 1.0 [102], tRNAs were predicted by tRNAscan-SE 1.23 [103], rRNAs were predicted by RNAmmer 1.2 [104], and functional annotation was finally performed based on homology searches against the RefSeq, TrEMBL, and Clusters of Orthologous Groups (COG) protein databases. Metabolic pathway prediction was performed on KAAS to assign KEGG Orthology (KO) numbers to each predicted CDS [105]. Annotations of the insertion sequences were conducted via the ISsaga web service [106]. Prophage regions were predicted using the PHAge Search Tool (PHAST) web server [107], and its results were confirmed by PCR runs with primers designed to detect phage attachment sites. CRISPR loci were searched for using the CRISPRFinder server [108].

The annotated genome was submitted to the GenomeRefine web service (http://genome.annotation.jp/genomerefine/), which assists with the refinement of annotations and registration at the DNA Data Bank of Japan (DDBJ).

2.2.4. *Comparative genome analysis*

The draft genome sequence of *L. suebicus* KCTC 3549^T was obtained from GenBank (accession no. BACO01000000). The genomic reads were downloaded from the DDBJ Sequence Read Archive for *L. oligofermentans* DSM 15707^T, *L. vaccinostercus* DSM 20634^T, and *L. vaccinostercus* DSM 15802 (accession nos. SRR1151187, SRR1151143, and ERR387466, respectively), which were assembled using the Platanus assembler. These genome sequences were annotated by MiGAP and KAAS in the same manner as *L. hokkaidonensis*. In addition, the genomic data were obtained for 13 representative species in the genus *Lactobacillus* from the NCBI Reference Sequence (RefSeq) database: *Lactobacillus acidophilus* NCFM (NC_006814), *Lactobacillus helveticus* DPC 4571 (NC_010080), *Lactobacillus delbrueckii* subsp. *bulgaricus* ATCC 11842 (NC_008054), *Lactobacillus gasseri* ATCC 33323 (NC_008530), *Lactobacillus reuteri* JCM 1112 (NC_010609), *Lactobacillus fermentum* IFO 3956 (NC_010610), *Lactobacillus buchneri* CD034 (NC_018610, NC_016035, NC_018611, NC_016034), *Lactobacillus brevis* ATCC 367 (NC_ 008497, NC_008498, NC_008499), *Lactobacillus casei* ATCC 334 (NC_008526, NC_008502, now labeled as *L. paracasei* ATCC 334), *Lactobacillus rhamnosus* GG

(NC_013198), *Lactobacillus plantarum* WCFS1 (NC_004567, NC_006375, NC_006376, NC_006377), *Lactobacillus sakei* subsp. *sakei* 23 K (NC_007576), and *Lactobacillus coryniformis* subsp. *coryniformis* CECT 5711 (NZ_AKFP00000000).

To compare the gene context, all-against-all BLASTP alignments were performed between *L. hokkaidonensis* LOOC260^T and each reference strain, and an ortholog table was constructed based on the bidirectional best hit among the BLAST results (Figure 2.7). BLAST alignments were obtained using the following thresholds: cut-off = E-value 0.0001 and ≥30% identity across ≥60% of the sequence length. Each row of the table represented a gene in *L. hokkaidonensis* LOOC260^T and its orthologous genes in the reference strains. For each row, the bit scores were divided by the maximum value. Therefore, the numbers in the cells denoted the normalized scores between 0 and 1. Each cell was colored a shade of red according to the normalized score with a deeper color corresponding to a higher score.

2.2.5. *Phylogenetic analysis*

A multiple alignment of 16S rRNA nucleotide sequences from 17 species included in the analysis was generated using MUSCLE [109]. The phylogenetic tree was constructed by Mega 5.0 using the neighbor-joining method with a bootstrap value of 1,000 [110].

2.2.6. *Data visualization*

The circular genome atlas shown in Figure 2.1 was produced using Circos software ver 0.66 [111] and in-house python scripts. The linear genome diagrams show in Figure 2.4 were generated using the GenomeDiagram module in BioPython [112] and they were adjusted manually.

2.2.7. *Availability of supporting data*

The complete genome sequence of *L. hokkaidonensis* LOOC260^T and its annotations were deposited at DDBJ/ENA/GenBank under accession numbers AP014680 (chromosome), AP014681 (plasmid pLOOC260-1), and AP014682 (plasmid pLOOC260-2). All of the sequencing data were deposited in the DDBJ Sequence Read Archive under accession numbers DRR024500 and DRR024501. The phylogenetic tree (Figure 2.6) and associated data matrix are available in TreeBASE database (Accession URL: http://purl.org/phylo/treebase/phylows/study/TB2: S17206).

2.3. Results and discussion

2.3.1. *Genome features of L. hokkaidonensis LOOC260*^{T}

Whole-genome sequencing was conducted with the PacBio single-molecule real-time (SMRT) sequencing system to determine the genome sequence of *L. hokkaidonensis* LOOC260^{T}. *De novo* assembly using the hierarchical genome assembly process (HGAP) method [113] generated seven contigs, which were further assembled and verified to finish the single complete genome. The genome of LOOC260^{T} comprises one circular chromosome (2,277,985 bp) and two circular plasmids designated as pLOOC260-1 (81,630 bp) and pLOOC260-2 (40,971 bp).

Two prophage regions were predicted, which are described in detail in the following section. No clustered regularly interspaced short palindromic repeat (CRISPR) loci were detected in the genome. The general genomic features of *L. hokkaidonensis* LOOC260^{T} and four other species in the *L. vaccinostercus* group are summarized in Table 2.1. Figure 2.1 shows the genome atlas of LOOC260^{T} as well as BLASTP alignment results with its four close relatives, as described above. Sharp transitions in the GCskew value were observed at both the predicted *oriC* site (0°) and its opposite site (176°). In particular, genes involved in metabolism (indicated in red) were densely encoded in the region from 300° to 360°. Several genes in this region were missing from all or some of the members of the *L. vaccinostercus* group, which may reflect the adaptation to specific ecological niches during the diversification of this group. Similar position-specific features have also been reported in *L. plantarum* [114] and *L. casei* [115], where they are considered to be lifestyle adaptation islands.

Table 2.1. Genome features of *L. hokkaidonensis* LOOC260^T and *L. vaccinostercus* group species.

Strain	Status	No. of sequences	Total bases	% GC	CDSs	rRNA operons	tRNA	INSD/SRA accession no.
L. hokkaidonensis LOOC260^T (Timothy grass silage)	Complete (Chromosome + 2 Plasmids)	3	2,277,985	38.2	2,194	4	56	AP014680#
			81,630	40.4	99	0	0	AP014681#
			40,971	39.4	51	0	0	AP014682#
L. oligofermentans DSM 15707^T (Modified atmosphere-packaged poultry products)	Scaffold	16	1,789,770	35.5	1,742	-	52	SRR1151187**
L. vaccinostercus DSM 20634^T (Cow dung)	Scaffold	88	2,551,457	43.5	2,471	-	52	SRR1151143**
L. vaccinostercus DSM 15802* [Acid-fermented condiment (tempoyak) in Malaysia]	Scaffold	129	2,558,791	43.5	2,506	-	53	ERR387466**
L. suebicus KCTC 3549^T (Apple mash)	Scaffold	143	2,656,936	39.0	2,583	-	55	BACO01

This study. * Formerly named *L. durianis*. ** SRA accession no.

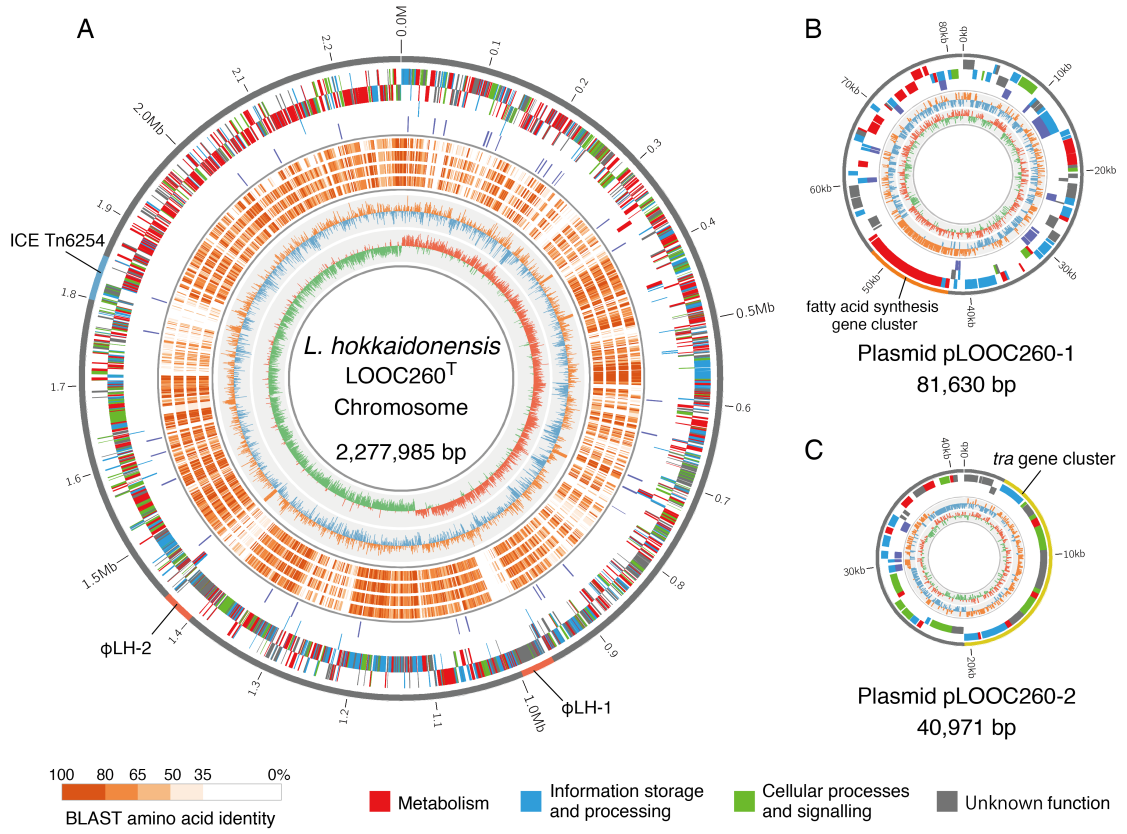


Figure 2.1. Genome atlas of *L. hokkaidonensis* LOOC260^T. A) Chromosome. The outer four circles (from outer to inner) represent CDSs on the forward strand, CDSs on the reverse strand, rRNAs (red) and tRNAs (blue), and insertion sequences/transposases, respectively. The next four circles (from outer to inner) represent the shared amino acid identities of the BLAST alignments with four closely related species: *L. oligofermentans* DSM 15707^T, *L. vaccinostercus* DSM 20634^T, *L. vaccinostercus* DSM 15802, and *L. suebicus* KCTC 3549^T, respectively. The inner two circles represent the GC content and GC skew. **B, C)** Plasmids pLOOC260-1 and pLOOC260-2. From outer to inner circles: CDSs on the forward strand, CDSs on the reverse strand, insertion sequences/transposases, GC content and GC skew. The CDSs are colored according to the main COGs functional classification categories: red, metabolism; blue, information storage and processing; red, cellular processes and signaling; gray, unknown function.

Bacterial genomes include various kinds of repetitive sequences such as multiple copies of ribosomal RNA operons and insertion sequences. These regions are generally difficult to reconstruct from relatively short sequencing reads, and thus *de novo* assembly often yields collapsed and/or fragmented contigs for such regions. To demonstrate the advantages of long-read assembly, contigs derived from MiSeq reads with the Platanus assembler were mapped to the reconstructed chromosomal sequence. The

yellow bands in Figure 2.2 represent the contigs, and most of the assembly gaps correspond to loci where rRNAs or transposase genes are encoded (outermost red and green bands, respectively). Other gaps were found within the large CDS regions, possibly encoding cell surface proteins with many repetitive sequences inside. Thus, the third-generation long-read sequencers can resolve such repetitive regions and show an excellent ability to reconstruct complete genome sequences. By contrast, it is difficult to reconstruct complete genomes using short-read sequencers alone; the draft genome reconstructed from MiSeq reads only consisted of 53 contigs (Table 2.2). However, draft genomes from short reads may be sufficient when conducting comparative analysis based on gene contents, as most of the CDSs except for repetitive genes could be reconstructed even from MiSeq reads. Indeed, 2,316 CDSs out of the 2,351 predicted CDSs in the complete genome of LOOC260^{T} were also found with perfect matches in the draft genome reconstructed using MiSeq reads and Platanus. The remaining 35 CDSs included those encoding transposase or those located in mobile elements. In this study, the complete genome derived from the PacBio sequencer with HGAP made it possible to capture a genome-wide perspective of mobile genetic elements like plasmids and prophages as described in the following subsection.

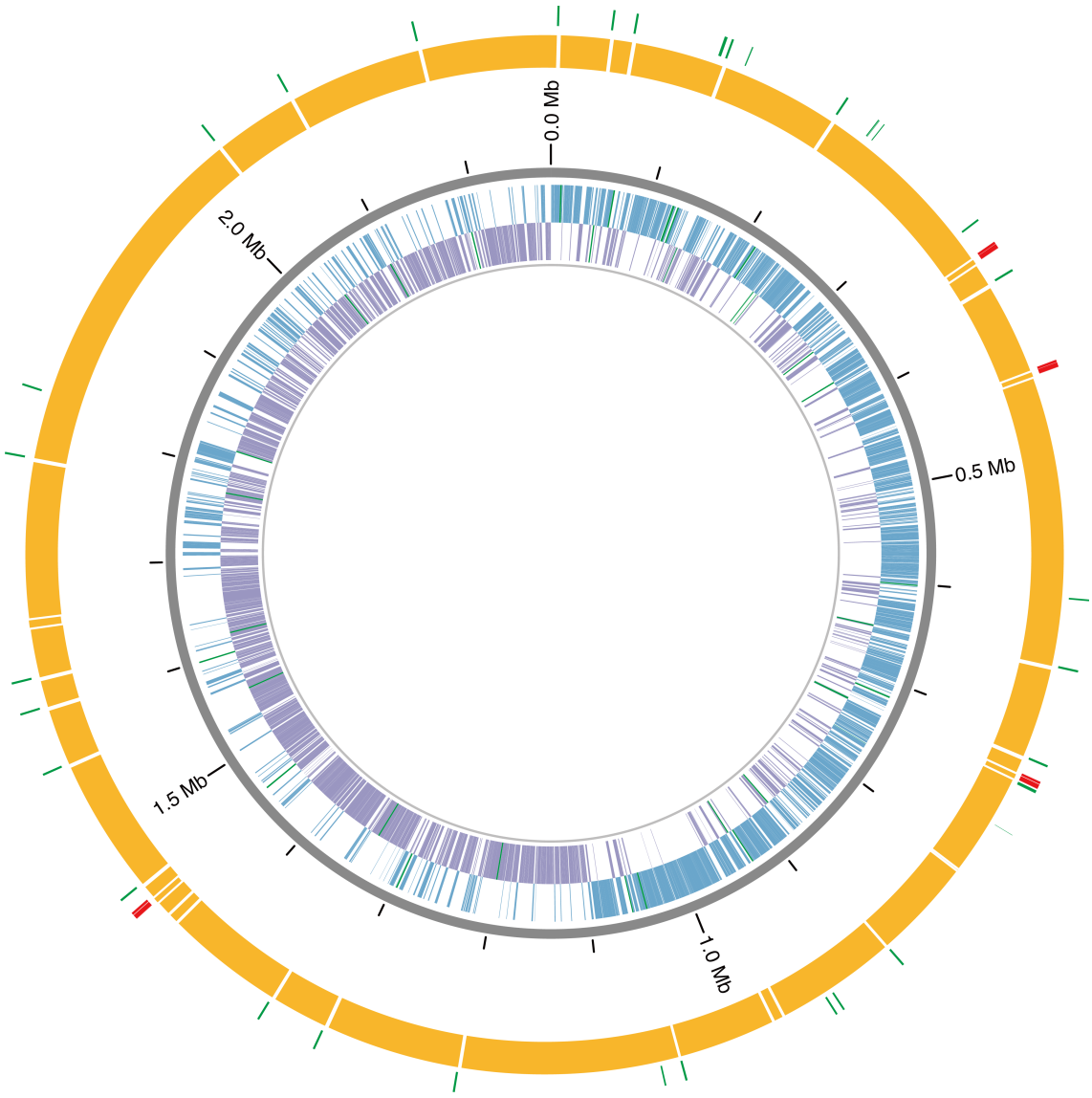


Figure 2.2. Comparison of contigs obtained from *de novo* assembly using MiSeq reads with the complete chromosomal sequence reconstructed using PacBio reads. The yellow bands represent contigs derived from MiSeq. Inner circles represent complete chromosomal sequences of *L. hokkaidonensis* LOOC260^{T} (CDSs on the forward strand and CDSs on the reverse strand from the inner to outer). The outermost bands represent rRNAs (red) and transposases genes (green).

Table 2.2. Comparison of assembly statistics.

Sequencing platform	Assembly method	No. of contigs	Total bases (bp)	N50 (bp)	Predicted CDSs	Predicted rRNAs	Predicted tRNA
PacBio SMRT	HGAP (draft)	7	2,513,068	1,771,111	–	–	–
PacBio SMRT	HGAP (finished)	3	2,400,586	–	2,344	12*	56
Illumina MiSeq	Platanus	53**	2,359,642	94,622	2,351	7***	56

* 4 copies of complete rRNA operons. ** Contigs shorter than 300 bp were eliminated. *** 5 copies of 5S rRNA and 2 partial sequences of 16S rRNA.

2.3.2. *Diverse mobile genetic elements*

Insertion sequences

In total, 59 ORFs, including partial ORFs and pseudogenes, were annotated as putative insertion sequences within the genome. In particular, three types of insertion sequence elements were annotated, with 13, 6, and 3 copies that shared almost 100% identity, and these new insertion sequence elements were registered in the ISfinder database [116] as ISLho1, ISLho2, and ISLho3, respectively. They shared 66% amino acid similarity with ISLre2 (*L. reuteri*), 75% with ISLrh2 (*L. rhamnosus*), and 60% with ISLre1 (*L. reuteri*), respectively.

Plasmids

The ratio of the mapped read number normalized against the sequence length for each replicon was approximately 1:1:4 (chromosome:pLOOC260-1:pLOOC260-2). Thus, the plasmid copy number in the cell was estimated as one for pLOOC260-1 and multiple for pLOOC260-2.

The first plasmid, pLOOC260-1, had a composite structure that comprised regions from several LAB species, such as *L. plantarum*, *L. casei*, *L. brevis*, and *L. coryniformis*, thereby indicating the occurrence of numerous rearrangements and recombination events during its evolution. The plasmid mobilization protein, Mob, gene was present, which probably facilitated the transmission of pLOOC260-1 in the presence of other conjugation mechanisms. Another interesting characteristic was the presence of a gene cluster related to fatty acid synthesis (LOOC260_200520–LOOC260_200630), which was absent from the chromosome. To the best of our knowledge, plasmid-encoded fatty acid synthesis

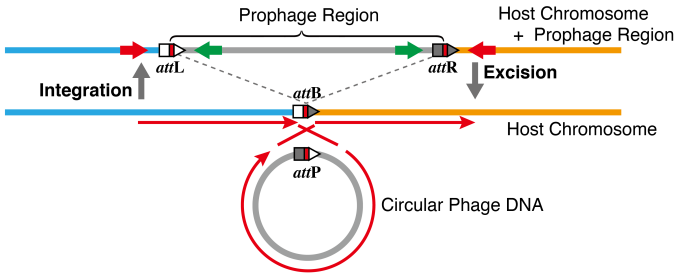
genes have not been reported previously in other LAB species.

The other plasmid, pLOOC260-2, was considered to be a conjugative plasmid. It possessed a *tra* conjugation gene cluster, which shared high similarity and colinearity with the plasmid pWCFS103 from *L. plantarum* WCFS1, for which conjugative transfer was demonstrated experimentally [117]. A similar gene organization in the *tra* region is also observed in several plant-associated LAB, such as *L. brevis* KB290, isolated from a Japanese fermented vegetable [118], *L. oryzae*, isolated from fermented rice grains [119,120], and *L. coryniformis*, frequently isolated from silage.

Prophages

Two prophage loci were predicted in the chromosome, ϕLH-1 (959–998 kb) and ϕLH-2 (1,400–1,437 kb). I also found 12-bp direct repeats (5'-TCACTCGCTTCA-3') flanking ϕLH-1 and 22-bp direct repeats (5'-ACT TAGAAAAATAAAAACGCGT-3') flanking ϕLH-2, which appeared to constitute the core regions of phage attachment sites (*att*R and *att*L). A contig obtained by *de novo* assembly contained a misassembled region that was presumably derived from an excised circular phage DNA, and thus spontaneous excision of the prophage must have occurred in a fraction of the cells. To confirm this prediction by PCR, two sets of primers were designed for each prophage so the fragments could be amplified only when the prophages were excised from the chromosome (Figure 2.3A, B). The expected PCR products were obtained, and the direct repeats located at the phage attachment sites were identified by sequencing the amplicons (Figure 2.3C). In the *L. vaccinostercus* group, these prophages are the first instances whose sequences have been determined and whose excision has been demonstrated.

A



B

	Name	Sequence	Position
→	Φ1_B_F	5′-GCCCATTGTAAATCCCTCCT-3′	(958,985–959,004)
←	Φ1_B_R	5′-ATTGCGTGTCTTAACGACCC-3′	(998,360–998,341)
→	Φ1_P_F	5′-CAGCACAGCAAGATTCAAGG-3′	(997,799–997,818)
←	Φ1_P_R	5′-ATCAGGATACTAGGACGACG-3′	(959,572–959,553)

	Name	Sequence	Position
→	Φ2_B_F	5′-CATTTAGGGACGAAGACCTC-3′	(1,399,994–1,400,013)
←	Φ2_B_R	5′-TGCTCAAGAGTTGCAGACAG-3′	(1,438,093–1,438,074)
→	Φ2_P_F	5′-GCGTGAATATATGCAGTGGG-3′	(1,436,894–1,436,913)
←	Φ2_P_R	5′-GCGTGAATATATGCAGTGGG-3′	(1,400,429–1,400,410)

C

ΦLH-1 attL (959,267–959,328)
AAGCGAATTGTCGTGAGTTCGAACCTCACTCGCTTCACTTTTCACAAGCCGTTATATACCTG
ΦLH-1 attR (998,034–998,095)
GGGATTTTTCGCACGTGTTCAACTGTCACTCGCTTCATTTTTTTACAAGCTGTCTATATATC
ΦLH-1 attB
AAGCGAATTGTCGTGAGTTCGAACCTCACTCGCTTCATTTTTTTACAAGCTGTCTATATATC
ΦLH-1 attP
GGGATTTTTCGCACGTGTTCAACTGTCACTCGCTTCACTTTTCACAAGCCGTTATATACCTG

ΦLH-2 attL (1,400,311–1,400,382)
AAAAGTAGAGTTTAGATCTTATATTACTTAGAAAAATAAAAACGCGTTACATTAATATGCTGTTAAATCAAC
ΦLH-2 attR (1,437,638–1,437,709)
TGTGACTCCAGATAATTCCCAAATCACTTAGAAAAATAAAAACGCGTCAACTCAAGAGCTGAATTGTCCATA
ΦLH-2 attB
AAAAGTAGAGTTTAGATCTTATATTACTTAGAAAAATAAAAACGCGTCAACTCAAGAGCTGAATTGTCCATA
ΦLH-2 attP
TGTGACTCCAGATAATTCCCAAATCACTTAGAAAAATAAAAACGCGTTACATTAATATGCTGTTAAATCAAC

Figure 2.3. Phage attachment sites. A) Schematic representation of the integration and excision of prophages. *att*P, *att*B, *att*L, and *att*R represent the phage attachment sites. Inward red arrows and outward green arrows indicate the PCR primers designed to amplify the *att*B and *att*P regions, respectively. **B)** Primer sequences for ϕLH-1 and ϕLH-2. **C)** Nucleotide sequences of the attachment sites for ϕLH-1 and ϕLH-2. Red letters represent the core sequences of the phage attachment sites. Blue and orange letters correspond to the left and right sequences of the prophage regions.

Integrated and conjugative elements

Integrated and conjugative elements (ICEs), sometimes known as conjugative transposons, are self-transmissible mobile genetic elements, which can be integrated into or excised from the host chromosome [121]. ICEs often contain accessory genes that confer advantages on their hosts, such as resistance to antibiotics, heavy metals, or phages [122]. ICEs have been reported frequently in *Streptococci* and *Enterococci*, and they are well characterized, but only two previous studies have described ICEs in the genus *Lactobacillus*: one in *L. paracasei* [123] and the other in *L. salivarius* [124]. No other ICEs are registered in the two ICEs/transposable elements databases: ICEberg [125] and Tn Number Registry [126].

I identified a putative ICE in the genome of LOOC260^{T} in the chromosome region 1,799–1,851 kbp (approximately 52 kbp), which was deposited as Tn6254 in the Tn Number Registry.

The amino acid sequences of the genes in the ICE region of LOOC260^{T} were compared against all the sequences in the NCBI non-redundant protein database, and similar gene organizations with high levels of sequence identity (>90%) were found in four species of plant-related LAB (Figure 2.4): *L. vini*

LMG 23202^{T} (isolated from grape must), *L. nodensis* JCM 14932^{T} (from rice bran), *L. paracasei* LPP49 (from cereal), and *L. coryniformis* (from cheese, silage, and kimchi). The level of shared nucleotide identity was also high between them. In particular, the 20-kbp upstream and 11-kbp downstream segments of Tn6254 were almost identical to the putative ICE from *L. vini* LMG 23202^{T}, but Tn6254 had more accessory genes, especially for heavy metal resistance, in the middle 21-kbp region. The integrase genes were adjacent to the 3'-end of the GMP-synthase gene, and direct repeats of 5'-GAGTGG GAATA-3' were identified at both the 3'-end of the GMP synthase gene and the 5'-end of the cell wall protein gene. The 3'-end of the GMP-synthase gene is reported to be an integration hotspot for genomic islands, and the consensus sequence of the direct repeats agreed with our findings [127]. However, in LOOC260^{T}, I found the same repeat sequence only at the integrase end and not at the opposite end because of the truncated 5'-end of the cell wall protein gene. Therefore, Tn6254 may no longer be capable of excision.

The shared sequence identities were high only within the strains described above. In particular, the four integrase genes shown in Figure 2.4 shared over 96% amino acid identity, whereas they exhibited lower identities (≤60%) with other known integrase genes. This suggests that these ICEs compose a single family and integrate themselves into the downstream region of the GMP-synthase. Heavy metal resistance genes are beneficial for plant-associated bacteria due to the fact that plants are exposed to metals in the soil, and may even absorb them. However, given their distinct ecological niches, it is unlikely that these ICEs were transferred directly between them. This suggests the existence of a large shared gene pool among plant-associated LAB.

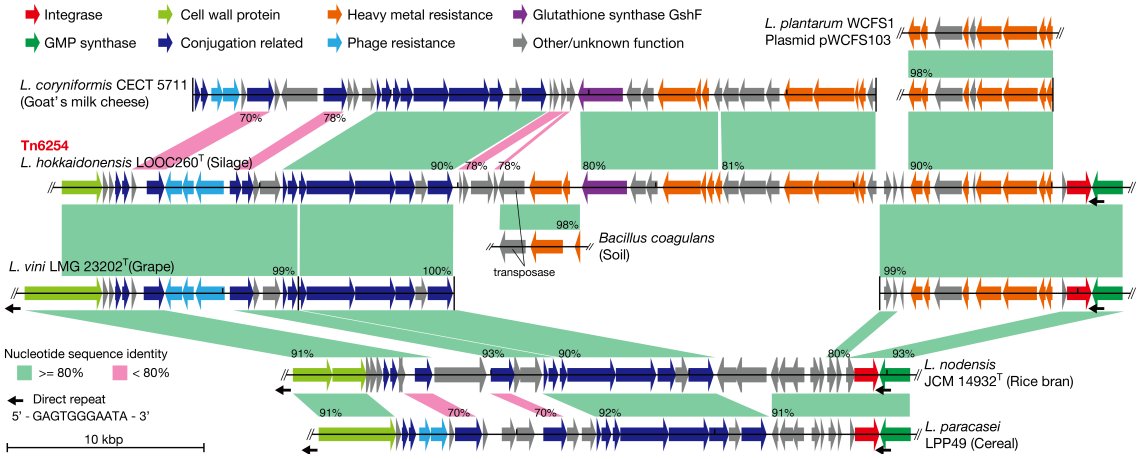


Figure 2.4. Comparisons of integrated and conjugative elements from *L. hokkaidonensis* and several species in the genus *Lactobacillus*. Green and red correspond to the nucleotide identity based on BLASTN alignments and the numbers indicate the identity. Small black arrows represent direct repeat sequences flanking the element.

2.3.3. *Cold adaptation strategy*

Cells exposed to low temperatures undergo significant physiological changes, such as decreases in membrane fluidity and stabilization of the secondary structures of nucleic acids, thereby resulting in less efficient transcription and translation [128]. In bacterial cell membranes, cold temperature induces fatty acid profile changes, such as the conversion of saturated fatty acids into unsaturated fatty acids and the preferential synthesis of shortchain, branched-chain, and/or anteiso fatty acids [129]. However, I found no distinctive characteristics related to the modification of fatty acid composition; I identified no genes involved in the synthesis of unusual fatty acids, such as unsaturated or branched-chain fatty acids, and I found that the number and order of the genes in the fatty acid biosynthesis gene cluster were identical to those in other species, except that they were encoded in the plasmid and not in the chromosome. Low temperatures also induce the production of several proteins such as cold shock protein A (CspA), which functions as an RNA chaperone, and RNA helicase DeaD, which prevents the formation of structured nucleic acids [130]. However, the numbers of these proteins differed slightly from those in the other 17 LAB strains included in the comparative analysis.

The cold stress response is also associated with different types of anti-stress mechanisms. Compatible solutes are chemical compounds, such as betaine and carnitine, that act as osmolytes and confer osmotic tolerance. They also facilitate psychrotolerance, although this physiological mechanism

still needs clarification [131]. The uptake and accumulation of compatible solutes in a cold-stressed environment, and the contributions of these solutes to psychrotolerance have been reported in several microorganisms, including *Listeria monocytogenes*, *Yersinia enterocolitica*, and *Bacillus subtilis* [131-133]. In *L. hokkaidonensis*, I found four transporters that were probably responsible for the uptake of these osmolytes: one BCCT family transporter (LOOC260_121750) and three ABC transporters (LOOC260_103390–103400, LOOC260_ 110220–110250, and LOOC260_117540–117560). The gene repertoire of these transporters was identical to that of *L. sakei*, a psychrotrophic LAB, in which the accumulation of compatible solutes is considered to be a key factor during acclimation to cold and saline environments [134]. Another notable feature was a bifunctional glutathione synthase encoded in the ICE region, GshF (LOOC260_118620), which allows glutathione to be synthesized via two-step ligation from its constituent amino acids [135]. Two key genes involved in the redox cycle of glutathione were also encoded: glutathione peroxidase (LOOC260_117530) and glutathione reductase (LOOC260_103410). Glutathione, which maintains cell redox homeostasis, also protects membrane lipids from the oxidative stress induced at cold temperatures [136,137]. In *L. hokkaidonensis*, GshF shared high similarity with that in *L. coryniformis*, which was a predominant isolate when screening for psychrotolerant LAB in Timothy grass silage (Figure 2.5), thereby indicating that glutathione may facilitate psychrotolerance in both species.

Bacterial defense systems that protect against cold environments involve a wide range of proteins, including those related to modifications of cell membrane lipids, transcription and translation mechanisms, and various stress proteins [129,130]; therefore, it is difficult to elucidate their direct evidence solely from the viewpoint of genomics. Hence, I will be conducting further investigations, including an expression study using whole transcriptome sequencing (RNA-seq). The changes of cellular status under stress conditions might be associated with metabolic shift that involves simultaneous gene regulation of multiple genes. Such global changes in cellular processes would be observed through the whole transcriptome analysis.

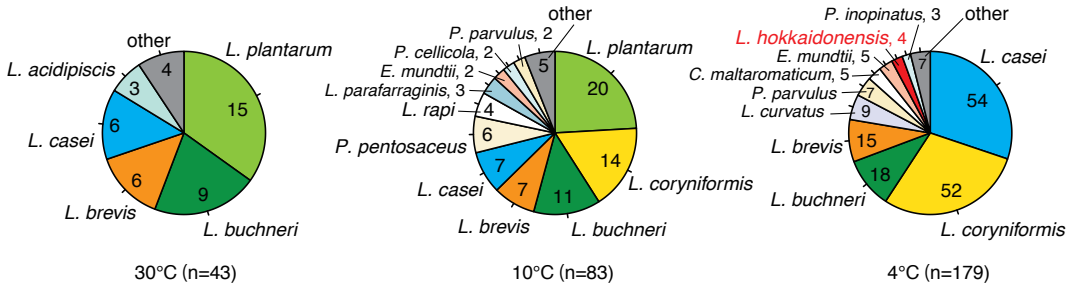


Figure 2.5. Distributions of species isolated from timothy grass silage stored in Hokkaido in the subarctic northern part of Japan during the winter season. Each strain was isolated after incubation on MRS agar plates for 10 days in anaerobic conditions at three different temperatures: 30°C, 10°C and 4°C. The 16S rRNA gene was sequenced for each isolate using the Sanger sequencing method. Species were identified by querying the sequences against the 16S rRNA sequence database downloaded from NCBI.

2.3.4. *Unique gene repertoire of the L. vaccinostercus group*

To clarify the characteristic gene features of *L. hokkaidonensis* and its close relatives, a comparative analysis was performed using four strains in the *L. vaccinostercus* group and 13 strains from representative LAB species. The phylogenetic tree of the 17 strains included in the analysis is shown in Figure 2.6. All-against-all bidirectional BLASTP alignments between *L. hokkaidonensis* LOOC260^{T} and each reference strain were conducted and an ortholog table was constructed based on the alignment results. Figure 2.7 shows an example of the table. In this analysis, a simple bidirectional best hit approach was adopted to obtain orthologous relationship between reference strains. Although the approach may not be the best way to find orthologous genes, it is a scalable approach to quickly find one-to-one relationship and to detect missing genes in reference strains. Although the number of strains in this study was small (17 strains), I used the same method for a much larger set in the recent studies.

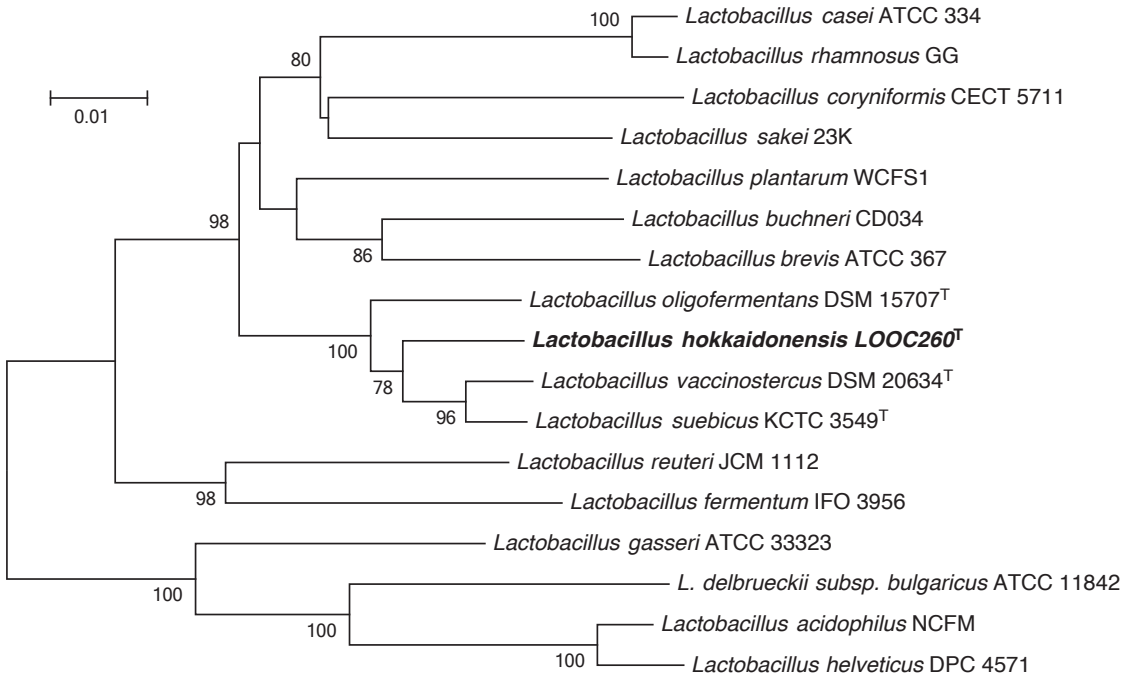


Figure 2.6. Neighbour-joining tree based on multiple alignment of the 16S rRNA nucleotide sequences from *Lactobacillus hokkaidonensis*, *L. vaccinostercus* group species (*L. vaccinostercus*, *L. suebicus*, *L. oligofermentans*), and 13 representative species in the genus *Lactobacillus*. The scale bar represents the number of substitution per site. Values at the nodes represent bootstrap values (1,000 replicates). Only values above 70% are shown.

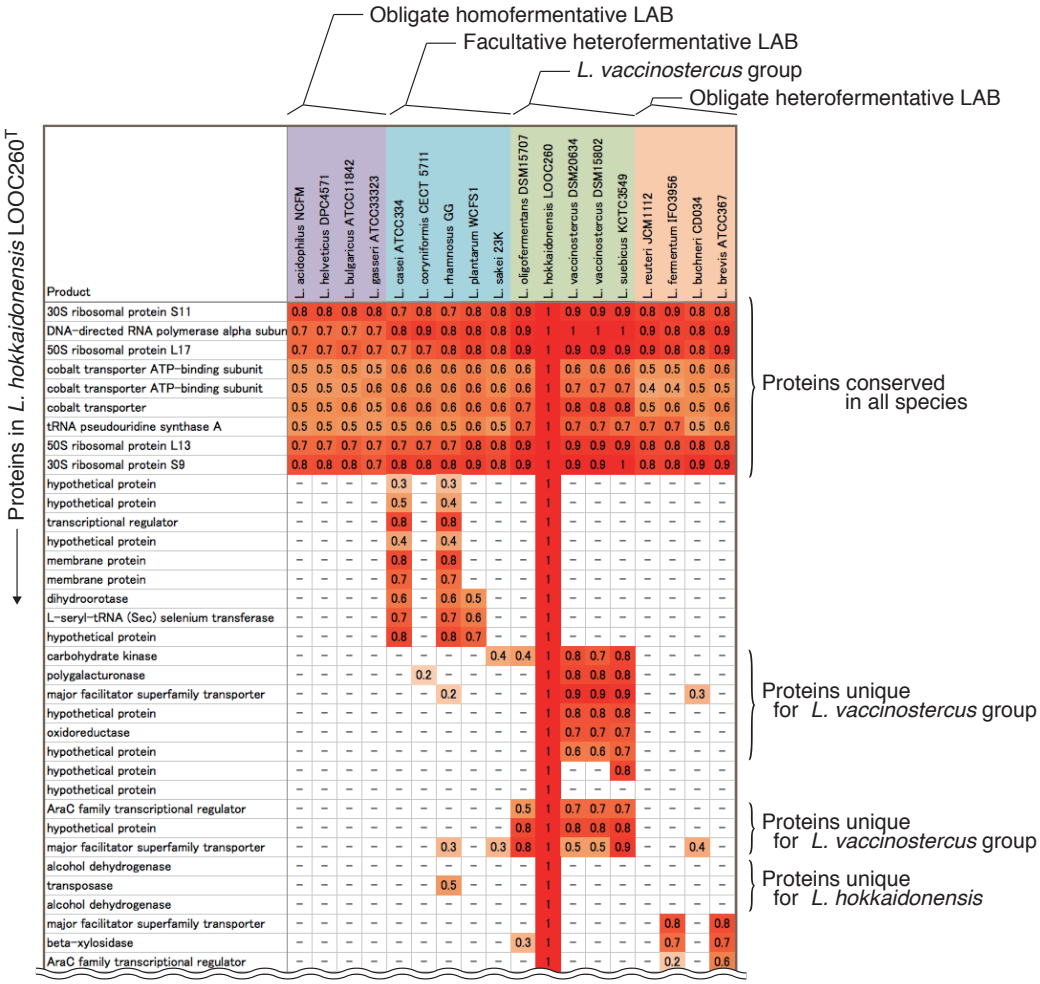


Figure 2.7. Ortholog table. Constructed based on the all-against-all BLASTP alignments between each two species. In the vertical direction, the proteins are shown in order of appearance in the genome of *L. hokkaidonensis* LOOC260^{T}. In the horizontal direction, the species included in the comparison are shown. For each row, the bit scores were normalized by dividing by the maximum value. The number of each cell represents the normalized score, and the cells are colored varying shades of red, according to their values, with a deeper color corresponding to a higher value.

Central metabolism

Similar to the well-characterized heterofermentative LAB, *L. buchneri* [138], all four species in the *L. vaccinostercus* group possessed phosphoketolase, a key enzyme in heterolactic fermentation, but they lacked two genes involved in the Embden-Meyerhof pathway: phosphofructokinase-1 and

fructose-bisphosphate aldolase. This was consistent with their classification as obligate heterofermentative LAB. They also possessed two key genes involved in the nonoxidative branch of the pentose phosphate pathway: transketolase and transaldolase. Both L and D-lactate dehydrogenase were encoded, which agrees with the phenotypic trait that both L-lactate and D-lactate are produced. In contrast to many of the obligate heterofermentative LAB, they lacked genes involved in the arginine deiminase pathway, which differentiates this group from the relatively closely related *L. reuteri* group. The reconstructed carbohydrate metabolism pathway is shown in Figure 2.8.

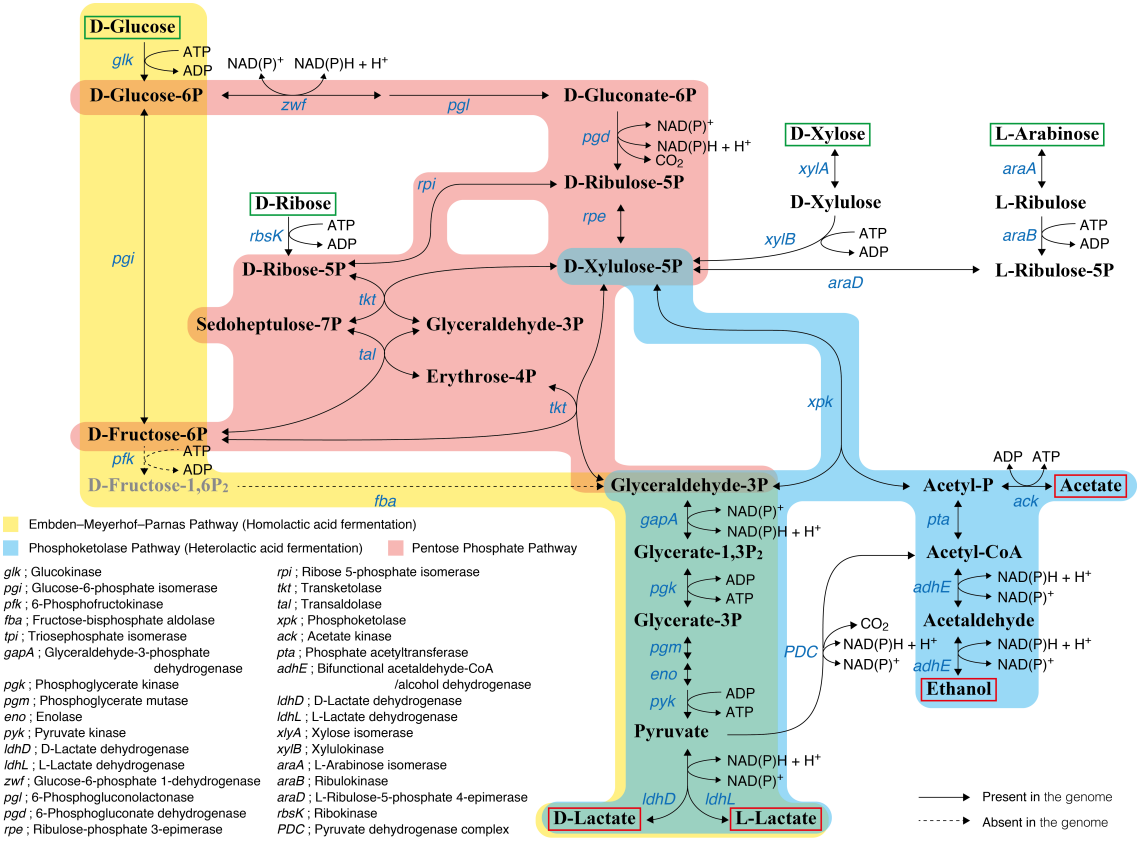


Figure 2.8. Reconstructed central carbohydrate metabolism pathway of *L. vaccinostercus* group species.

The species in the *L. vaccinostercus* group can assimilate pentoses, such as L-arabinose, D-ribose, and D-xylose, more rapidly than D-glucose, thereby indicating a preference for pentoses over hexoses [90,96,139]. The weak capacity for glucose utilization may be attributed to the cellular redox imbalance caused

by insufficient regeneration of NAD(P)$^+$ because the *L. vaccinostercus* growth rate on glucose is accelerated by adding electron acceptors, such as aldehydes and ketones, to the medium [139,140]. These characteristics are similar to *Fructobacillus* species, which lack the *adhE* (bifunctional alcohol/acetaldehyde dehydrogenase) gene for regenerating NAD(P)$^+$ in the latter stage of heterolactic fermentation [141]. By contrast, members of the *L. vaccinostercus* group possess *adhE*, which suggests that another mechanism is active.

As a starter culture for silage fermentation, the ability to assimilate pentoses is advantageous when utilizing substrates derived from plant cell walls. Hemicellulose is one of the major components of the plant cell wall, which is composed of a branched heteropolymer of saccharides [65]. During the ensiling process, hemicellulose is partially hydrolyzed to yield pentoses, such as xylose and arabinose, which are then fermented into lactic and acetic acid via the phosphoketolase pathway [142]. In addition, acetic acid acts as an effective inhibitor that prevents the growth of aerobic spoilage microorganisms, such as yeasts and molds, thereby improving stability against aerobic deterioration after silos are opened for feeding [143]. In addition to the genes necessary to ferment pentoses, the presence of several copies of β-xylosidase genes in *L. hokkaidonensis* (LOOC260_101610, LOOC260_101740, and LOOC260_105960) indicates the ability to utilize xylooligosaccharide.

NADPH generation

Unique mechanisms were found for NADPH generation in the *L. vaccinostercus* group LAB. *L. hokkaidonensis*, *L. vaccinostercus*, and *L. suebicus* possessed membranebound NAD(P) transhydrogenase PntAB, which mediates the transfer of a hydrogen from NADH to NADP$^+$ to produce NADPH using the electrochemical proton gradient [144]. In addition, *L. vaccinostercus* and *L. suebicus* possessed NADP-dependent glyceraldehyde-3-phosphate dehydrogenase, GapN, which catalyzes the one-step conversion of glyceraldehyde-3-phosphate to 3-phosphoglycerate, with the concomitant reduction of NADP$^+$ to NADPH [145]. In conventional glycolysis, glyceraldehyde-3-phosphate is converted into 3-phosphoglycerate via a two-step reaction, which is accompanied by the formation of NADH and ATP.

The major cellular source of NADPH is considered to be the oxidative branch of the pentose phosphate pathway, where hexoses are decarboxylated into a C5-moiety. However, pentoses are assimilated without passing through this branch; thus, these enzymes may provide an alternative route for generating NADPH. Analogously, GapN in *Streptococcus mutans*, which lacks the oxidative part of the pentose phosphate pathway, has been suggested to participate in NADPH generation [146]. NADPH mainly functions as an electron donor in anabolic reactions, whereas NAD$^+$ mainly functions as an electron acceptor in catabolic reactions. Therefore, both PntAB and GapN are favorable, particularly in the

biosynthetic process because they produce a higher $NADPH/NADP^+$ ratio and a lower $NADH/NAD^+$ ratio.

With the exception of the meat-borne *L. oligofermentans*, the *L. vaccinostercus* group LAB members encode a relatively high number of genes for amino acid biosynthesis. These NADPH generation systems may support the diverse biosynthetic abilities of *L. hokkaidonensis* and its close relatives and may reflect the optimized utilization of pentoses as growth substrates.

2.4. Conclusions

In this study, the complete genome of *L. hokkaidonensis* $LOOC260^T$ was successfully reconstructed by whole-genome sequencing using the PacBio SMRT sequencing system and *de novo* assembly based on the HGAP method. The complete genome of *L. hokkaidonensis* $LOOC260^T$ contained various previously unreported mobile genetic elements, which included three new types of insertion sequences, two prophage loci, one ICE, and two plasmids, one of which was considered to be a conjugative plasmid. ICE contained many genes related to heavy metal resistance and shared several components with other plant-associated LAB. The ICE may have mediated the dissemination of genes that contributed to niche adaptation in plant-associated LAB species. The comparative genome analysis also provided insights into the characteristic gene repertoire of this group, such as preferential pentose assimilation. Although this study could not obtain direct evidence of psychrotolerance, I detected possible factors that may contribute to psychrotolerance in this species, such as the uptake of compatible solutes and the synthesis of glutathione. These findings merit further investigations, and the genomic information obtained in this study should facilitate the development of an appropriate silage inoculant for use in cold regions.

Chapter 3

3. Comparative genomics of *Fructobacillus* spp. and *Leuconostoc* spp.

3.1. Introduction

Lactic acid bacteria (LAB) are found in a variety of environments, including dairy products, fermented food or silage, and gastrointestinal tracts of animals. Their broad habitats exhibit different stress conditions and nutrients, forcing the microbe to develop specific physiological and biochemical characteristics, such as proteolytic and lipolytic activities to obtain nutrients from milk [147], tolerance to phytoalexins in plants [148], or tolerance to bile salts to survive in the gastrointestinal tracts [149]. *Fructobacillus* spp. in the family *Leuconostocaceae* are found in fructose-rich environments such as flowers, (fermented) fruits, or bee guts, and are characterized as fructophilic lactic acid bacteria (FLAB) [150-152].

The genus *Fructobacillus* is comprised of five species: *Fructobacillus fructosus* (type species), *F. durionis*, *F. ficulneus*, *F. pseudoficulneus* and *F. tropaeoli* [152,153]. Four of the five species formerly belonged to the genus *Leuconostoc*, but were later reclassified as members of a novel genus, *Fructobacillus*, based on their phylogenetic position, morphology, and biochemical characteristics [79]. *Fructobacillus* is distinguished from *Leuconostoc* by the preference for fructose over glucose as the carbon source and the need for an electron acceptor (e.g. pyruvate or oxygen) during glucose assimilation. *Fructobacillus* is further differentiated from *Leuconostoc* by the production of acetic acid instead of ethanol when glucose is metabolized. The previous study revealed that *Fructobacillus* lacked the bifunctional acetaldehyde/alcohol dehydrogenase gene (*adhE*) and its enzyme activities [141], which might be relevant to the acetic acid production from glucose. They are the only obligately heterofermentative LAB without *adhE* to date, suggesting that niche-specific evolution occurred at the genome level. Recent comparative genomic studies also revealed niche-specific evolution of several LAB, including vaginal lactobacilli and strains used as dairy starter cultures [154-156].

This is the first study to compare the metabolic properties of the draft genome sequences of five *Fructobacillus* spp. with those of *Leuconostoc* spp., with a special focus on fructose-rich niches. Results obtained confirm the general trend of reductive evolution, especially metabolic simplification based on

sugar availability.

3.2. Methods

3.2.1. *Bacterial strains and DNA isolation*

Fructobacillus fructosus NRIC 1058^T, *F. ficulneus* JCM 12225^T, *F. pseudoficulneus* DSM 15468^T and *F. tropaeoli* F214-1^T were cultured in FYP broth (l^{-1}: 10 g Dfructose, 10 g yeast extract, 5 g polypeptone, 2 g sodium acetate, 0.5 g Tween 80, 0.2 g $MgSO_4 \cdot 7H_2O$, 0.01 g $MnSO_4 \cdot 4H_2O$, 0.01g $FeSO_4 \cdot 7H_2O$, 0.01g NaCl; pH 6.8) at 30 °C for 24 h. Genomic DNA was isolated by the method of a combination of phenol/chloroform and glass beads as described by Endo and Okada [157].

3.2.2. *Genome sequences used in this study*

Whole-genome sequencing was conducted by Illumina Genome Analyzer II system, with insert length of about 500 bp. Total 6,060,140, 1,904,646, 2,474,758 and 13,680,640 reads with average lengths of 60 to 91 bp were obtained from *F. fructosus* NRIC 1058^T, *F. ficulneus* JCM 12225^T, *F. pseudoficulneus* DSM 15468^T and *F. tropaeoli* F214-1^T, respectively. *De novo* assembly using the Velvet Assembler for short reads with parameters optimized by the VelvetOptimizer (Version 1.2.10) [158] resulted in 57, 28, 15 and 101 contigs each (Length: 1,489,862, 1,552,198, 1,413,733 and 1,686,944 bp; N_{50}: 89,458, 226,528, 283,981 and 226,443 bp). The *k*-mer sizes for the strains were 81, 45, 51, 63 bp each. The genome was annotated using the Microbial Genome Annotation Pipeline (MiGAP) [101] with manual verification. In the pipeline, protein coding sequences (CDSs) were predicted by MetaGeneAnnotator 1.0 [102], tRNAs were predicted by tRNAscan-SE 1.23 [103], rRNAs were predicted by RNAmmer 1.2 [104], and functional annotation was finally performed based on homology searches against the RefSeq, TrEMBL, and Clusters of Orthologous Groups (COG) [159] protein databases.

Draft genome sequence of *Fructobacillus durionis* DSM 19113^T was obtained from the JGI Genome Portal (http://genome.jgi.doe.gov/) [160] and annotated using MiGAP in the same way as other *Fructobacillus* spp. Annotated genome sequences for nine of the twelve *Leuconostoc* species were obtained from the GenBank databases at NCBI. Of *Leuconostoc* spp., genomic data of *Leuconostoc holzapfelii*, *Leuconostoc miyukkimchii* and *Leuconostoc palmae* were not available at the time of analysis (December 2014) and were not included in the present study. When multiple strains were available for a single species, the most complete one was chosen. GenBank accession numbers of the strains used are

listed in Table 3.1.

3.2.3. *Quality assessment of the genomic data*

The completeness and contamination of the genomic data were assessed by CheckM (Version 1.0.4) [92], which inspects the existence of gene markers specific to the *Leuconostocaceae* family, a superordinate taxon of *Fructobacillus* and *Leuconostoc*.

3.2.4. *Comparative genome analysis and statistical analysis*

To estimate the size of conserved genes, all protein sequences were grouped into orthologous clusters by GET_HOMOLOGUES software (version 1.3) based on the all-against-all bidirectional BLAST alignment and the MCL graph-based algorithm [161]. The conserved genes are defined as gene clusters that are present in all analyzed genomes (please note the difference from the definition of specific genes). The rarefaction curves for conserved and total genes were drawn by 100-time iterations of adding genomes one by one in a random order. From this analysis, two genomes (*L. fallax* and *L. inhae*) were excluded to avoid underestimation of the size of conserved genes, since they contained many frameshifted genes, probably due to the high error rate at homopolymer sites of Roche 454 sequencing technology.

For functional comparison of the gene contents between *Fructobacillus* spp. and *Leuconostoc* spp., CDS predicted in each strain were assigned to Cluster of Orthologous Groups (COG) functional classification using the COGNITOR software [159]. Metabolic pathway in each strain was also predicted using KEGG Automatic Annotation Server (KAAS) by assigning KEGG Orthology (KO) numbers to each predicted CDS [105]. The numbers of genes assigned to each COG functional category were summarized in Table 3.2. In the present study, *Fructobacillus*-specific genes were defined as those conserved in four or more *Fructobacillus* spp. (out of five) and in two or less *Leuconostoc* spp. (out of nine). *Leuconostoc*-specific genes were defined as those conserved in seven or more *Leuconostoc* spp. and one or less *Fructobacillus* spp. (Table 3.4).

The Mann–Whitney U test was applied to compare genome features and gene contents of *Fructobacillus* spp. and *Leuconostoc* spp. The p value of 0.05 was considered statistically significant. Statistical analysis was performed using IBM SPSS Statistics for Windows (Version 21.0. Armonk, NY: IBM Corp.).

3.2.5. *Phylogenetic analysis*

Orthologous clusters that were conserved among all *Fructobacillus* spp., all *Leuconostoc* spp. and *Lactobacillus delbrueckii* subsp. *bulgaricus* ATCC 11842^T (as an outgroup) were determined by GET_HOMOLOGUES as described above. For phylogenetic reconstruction, 233 orthologs that appeared exactly once in each genome were selected. The amino acid sequences within each cluster were aligned using MUSCLE (version 3.8.31) [109]. Poorly aligned or divergent regions were trimmed using Gblocks [162], and conserved regions were then concatenated using FASconCAT-G [163]. A partitioned maximum likelihood analysis was performed to construct the phylogenetic tree with RAxML (version 8.1.22) [164] using the best fit evolutionary models predicted for each alignment by ProtTest [165]. The number of bootstrapping was 1,000 replicates.

3.2.6. *Polysaccharides production and reaction to oxygen*

Polysaccharides production from sucrose were determined by the methods as described Endo and Okada [166]. Briefly, the strains were inoculated on agar medium containing sucrose as sole carbon source and incubated aerobically at 30 °C for 48 h.

To investigate the reaction to oxygen on growth, the cells were streaked onto GYP agar, which contained D-glucose as the sole carbon source, and cultured under anaerobic and aerobic conditions at 30 °C for 48 h [79,150]. The anaerobic conditions were provided by means of a gas generating kit (AnaeroPack, Mitsubishi Gas Chemical, Japan). These studies were conducted for the type strains of five *Fructobacillus* species, *Leuconostoc mesenteroides* subsp. *mesenteroides* NRIC 1541^T, *Leuconostoc citreum* NRIC 1776^T and *Leuconostoc fallax* NRIC 0210^T.

3.2.7. *Data deposition*

Annotated draft genome sequences of *F. fructosus* NRIC 1058^T, *F. ficulneus* JCM 12225^T, *F. pseudoficulneus* DSM 15468^T and *F. tropaeoli* F214-1^T were deposited to the DDBJ/EMBL/GenBank International Nucleotide Sequence Database with accession numbers BBXR01000000, BBXQ01000000, BBXS01000000 and BBXT01000000, respectively. Unassembled raw sequence data were also deposited to the database with accession number DRA004155. The phylogenetic tree and associated data matrix for Fig. 3.6 are available at TreeBASE (Accession URL: http://purl.org/phylo/treebase/phylows/study/TB2:S18090).

3.3. Results and discussion

3.3.1. *General genome features of Fructobacillus spp. and Leuconostoc spp.*

Draft genome sequences of four *Fructobacillus* spp. were determined by the Illumina Genome Analyzer II system. The sequence coverage of *F. fructosus* NRIC 1058^T, *F. ficulneus* JCM 12225^T, *F. pseudoficulneus* DSM 15468^T and *F. tropaeoli* F214-1^T were 329-, 55-, 90-, and 513-fold, respectively. Genome sequences of nine *Leuconostoc* spp. and *Fructobacillus durionis* were obtained from public databases (see Methods). The genome features of the strains used in the present study are summarized in Table 3.1. The genome sizes of *Fructobacillus* ranged from 1.33 to 1.69 Mbp (median ± SD, 1.49 ± 0.30 Mbp) and are significantly smaller than those of *Leuconostoc* ($p < 0.001$), 1.69 to 2.30 Mbp (median ± SD, 1.94 ± 0.21) (Figure 3.1A). Accordingly, *Fructobacillus* strains contain significantly smaller numbers of CDSs than *Leuconostoc* strains (median ± SD, 1387 ± 132 vs 1980 ± 323, $p < 0.001$) (Figure 3.1B). The DNA G + C contents of both species are also significantly different ($p < 0.001$): median ± SD is 44.4 % ± 0.30 % in *Fructobacillus* and 38.1 % ± 2.05 % in *Leuconostoc* (Figure 3.1C). These distinct genomic features strongly support the reclassification of *Fructobacillus* spp. from the genus *Leuconostoc*. The difference in G + C content is more prominent at the third positions of codons (GC3): 46.0 % ± 1.02 % in *Fructobacillus* and 30.9 % ± 4.12 % in *Leuconostoc*. The third nucleotides of codons are more likely to change rapidly than the first or the second ones because mutations at the third positions are often silent. The difference in GC3 may indicate different evolutional directions of the two genera. Similarly, *Lactobacillus delbrueckii*, also known to be in the ongoing process of genome reduction, exhibits the higher GC3 value [155]. The trend in *Fructobacillus* and *L. delbrueckii* is opposite to the general tendency that bacterial species with smaller genomes exhibit lower G + C contents [211]. However, whether this trend can be generalized to other LAB is not clear.

Table 3.1. General genome characteristics of the strains analyzed.

Strains	Status[1]	Source	INSD/SRA Accession no.	Size (Mbp)	No. of CDS	%G+C	%GC3	Completeness[3]	Contamination[3]
Fructobacillus fructosus NRIC 1058^T	D	Flower	BBXR01000000	1.49	1437	44.6	46.4	93.62	0
Fructobacillus durionis DSM 19113^T	D	Fermented fruit	JGI[2]	1.33	1221	44.7	47.4	94.98	0.57
Fructobacillus ficulneus JCM 12225^T	D	Fig	BBXQ01000000	1.55	1397	43.9	44.6	92.79	0.48
Fructobacillus pseudoficulneus DSM 15468^T	D	Fig	BBXS01000000	1.41	1312	44.5	45.9	95.14	0.48
Fructobacillus tropaeoli F214-1^T	D	Flower	BBXT01000000	1.68	1572	44.2	45.7	94.98	0.24
Leuconostoc mesenteroides ATCC 8293^T	C	Fermenting olives	CP000414-15	2.08	2045	37.7	30.1	100	0
Leuconostoc carnosum JB16	C	Kimchi	CP003851-55	1.77	1696	37.1	27.9	99.04	0.6
Leuconostoc citreum KM20	C	Kimchi	DQ489736-40	1.9	1849	38.9	31.3	99.52	0
Leuconostoc fallax KCTC 3537^T	D	Sauerkraut	AEIZ01000000	1.64	1882	37.5	29.2	97.3	1.16
Leuconostoc gelidum JB7	C	Kimchi	CP003839	1.89	1818	36.7	27.6	99.04	0.24
Leuconostoc inhae KCTC 3774^T	D	Kimchi	AEMJ01000000	2.3	2790	36.4	28.6	95.59	5.38
Leuconostoc kimchii IMSNU 11154^T	C	Kimchi	CP001753-58	2.1	2097	37.9	30.1	99.52	0
Leuconostoc lactis KACC 91922	D	Kimchi	JMEA01000000	1.69	2076	43.4	41.1	99.04	0.57
Leuconostoc pseudomesenteroides 1159	D	Cheese starter	JAUI01000000	2.04	1634	39	32.5	99.04	0.16

[1] Status: D, draft genome; C, complete genome.

[2] Obtained from Integrated Microbial Genomes (IMG) database at the Department of Energy Joint Genome Institute (http://genome.jgi-psf.org)

[3] Calculated using CheckM (version 1.04)

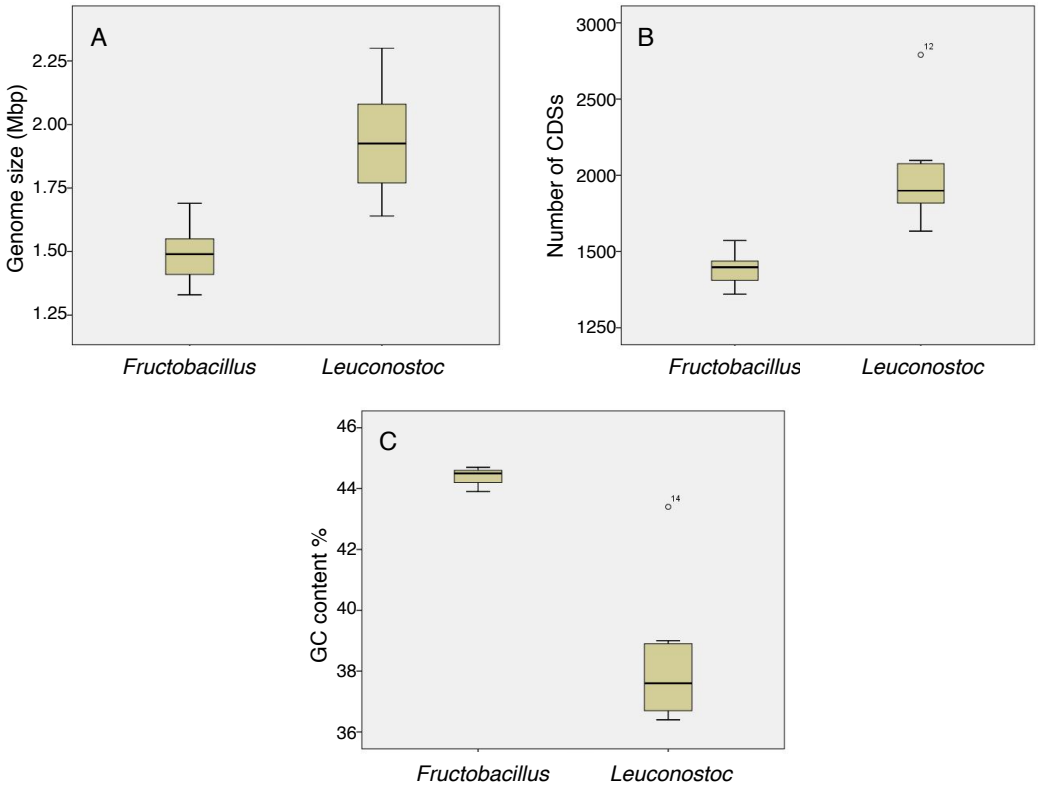


Figure 3.1. Genome sizes (A), number of CDSs (B) and GC contents (C) in *Fructobacillus* spp. and *Leuconostoc* spp. The line in the box represents the median, with lower line in the 25% border and the upper line the 75% border. The end of the upper vertical line represents the maximum data value, outliers not considered. The end of the lower vertical line represents the lowest value, outliers not considered. The separate dots indicate outliers.

Since most of the genomes analyzed in this study were in draft status, quality assessment of the genomes was conducted using CheckM. The average completeness values for *Fructobacillus* and *Leuconostoc* genomes were 94.3 and 98.7 %, respectively (Table 3.1). Except for the genome of *L. inhae*, which exhibited the contamination value of 5.4 %, all genomes satisfied the criteria required to be considered a near-complete genome with low contamination (≥90 % completeness value and ≤ 5 % contamination value) [92]. The lower completeness values for *Fructobacillus* genomes might be attributable to insufficiency of the reference gene markers used by CheckM, for which the genomic data of *Fructobacillus* spp. were not reflected at the time of writing this paper (December 2014), rather than the

lower quality of these genomes. In addition, the lower completeness may indicate specific gene losses in the genus *Fructobacillus* since the closer investigation of CheckM results showed that seven gene markers were consistently absent among five *Fructobacillus* genomes while on average, 14.6 markers were absent out of 463 *Leuconostocaceae*-specific gene markers.

3.3.2. *Conserved genes in Fructobacillus spp. and Leuconostoc spp.*

The numbers of conserved genes in the nine genomes of *Leuconostoc* and five genomes of *Fructobacillus* were estimated as 1,026 and 862, respectively. They account for 52 % and 62 % of average CDS numbers of each genus (Figure 3.2A). The difference in the average CDS numbers reflects their genomic history including ecological differences between the two genera. A previous study also reported 1,162 conserved genes in three genomes of *Leuconostoc* species [167]. The smaller number and the higher ratio of fully conserved genes in *Fructobacillus* spp. are probably due to a less complex and consistent habitat with specific sugars only, such as fructose. It is a major carbohydrate found in habitats of *Fructobacillus* spp., e.g. flowers, fruits and associated insects. On the other hand, *Leuconostoc* spp., that are usually seen in wide variety of habitats, including gut of animals, dairy products, plant surfaces, or fermented foods and soils, possess a larger number of conserved genes. Figure 3.2B shows the distribution of gene clusters in two genera. The frontmost peak (721 gene clusters) represents conserved genes that are shared by both *Leuconostoc* and *Fructobacillus* spp. Genus-specific conserved genes are indicated as leftmost and right peaks in Figure 3.2B. The leftmost peak (159 gene clusters) represents genes that are present in all *Leuconostoc* genomes, but absent in all *Fructobacillus* genomes, and the right peak (24 gene clusters) represents vice versa. The much smaller peak of the right compared to that of the left indicates that *Fructobacillus* spp. have lost more genes or have acquired less genes than *Leuconostoc* spp. during diversification after they separated into two groups. In addition, the number of gene clusters located near the center of the figure was small, which indicates that the exchange of genes between the two genera is not frequent and that they share distinct gene pools. This supports the validity of the classification of *Fructobacillus* as a distinct genus [79].

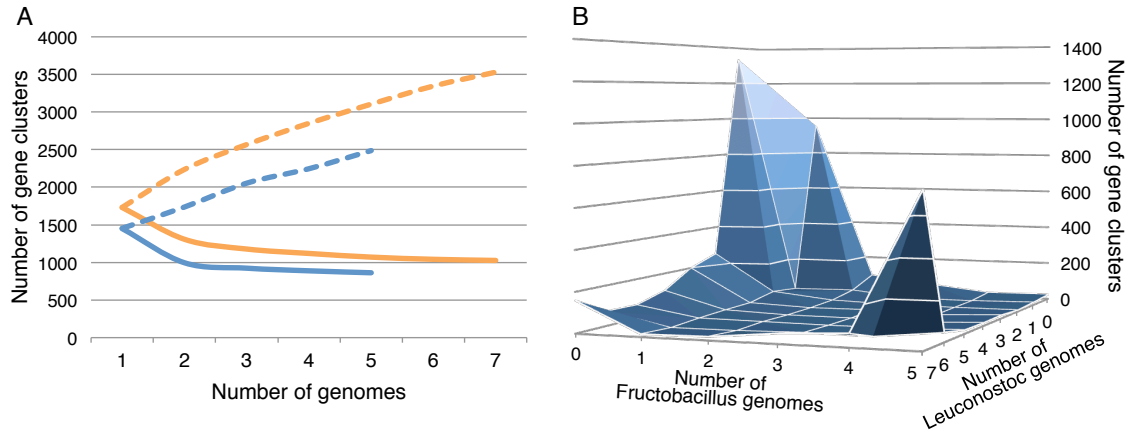


Figure 3.2. Core- and pan-genome of *Fructobacillus* and *Leuconostoc*. A) Estimation of the numbers of core- and pan-genome for *Fructobacillus* (blue) and *Leuconostoc* (orange). Solid lines represent core- and dashed lines represent pan-genomes as a function of the number of genomes added. The medium of 100 random permutations of the genome order is presented. **B)** Distribution of gene clusters present in *Fructobacillus* and *Leuconostoc*. Horizontal axes represent the numbers of genomes in each genus. Vertical axes show the numbers of gene clusters present in the given number of genomes.

3.3.3. *Comparison of gene contents between Fructobacillus spp. and Leuconostoc spp.*

The identified genes were associated with COG functional categories by COGNITOR software at the NCBI. The numbers of genes assigned in each COG category were summarized in Figure 3.3 and Table 3.2. *Fructobacillus* spp. have less genes for carbohydrate transport and metabolism compared to *Leuconostoc* spp. (Class G in Figure 3.3): Class G ranked 9th largest in *Fructobacillus* whereas it ranked 3rd in *Leuconostoc*. Similarly, the number of genes in Class C (energy production and conversion) was significantly less in *Fructobacillus* spp. than in *Leuconostoc* spp., suggesting that energy systems in *Fructobacillus* spp. are much simpler than those in *Leuconostoc* spp. The smaller number of CDS and conserved genes in *Fructobacillus* spp. could have resulted from metabolic reduction caused by scarce availability of carbohydrates other than fructose. Oppositely, the numbers of genes assigned in Class D (cell cycle, cell division and chromosome partitioning), Class J (translation, ribosomal structure and biogenesis), Class L (replication, recombination and repair) and Class U (intracellular trafficking, secretion and vesicular transport) were comparable between *Fructobacillus* spp. than in *Leuconostoc* spp. The conservation of genes in these classes against the genome reduction may indicate that their functions

are essential for re-production, and the class names roughly correspond to housekeeping mechanisms.

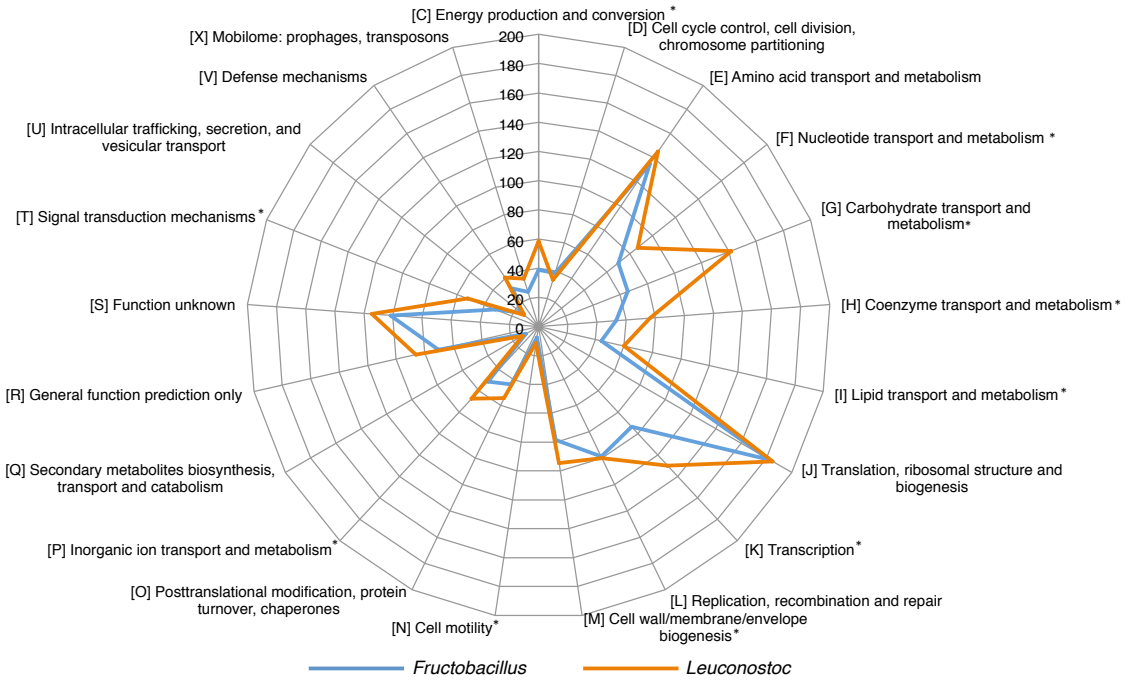


Figure 3.3. Comparison of gene content profiles obtained for the genera *Fructobacillus* and *Leuconostoc*. The values represent the numbers of genes assigned to specific COG categories. The Mann-Whitney *U* test was done to compare *Fructobacillus* spp. and *Leuconostoc* spp., and significant differences ($P<0.05$) are denoted with an asterisk (*).

Table 3.2. Gene content profiles obtained for *Fructobacillus* spp. and *Leuconostoc* spp.

	F. fructosus NRIC 1058^T	*F. durionis* DSM 19113^T	*F. ficulneus* JCM 12225^T	*F. pseudoficulneus* DSM 15468^T	*F. tropaeoli* F214-1^T	*L. mesenteroides* ATCC 8293^T	*L. carnosum* JB16	*L. citreum* KM20	*L. fallax* KCTC 3537^T	*L. gelidum* JB7	*L. inhae* KCTC 3774^T	*L. kimchii* IMSNU 11154^T	*L. lactis* KACC 91922	*L. pseudomesenteroides* 1159
[C] Energy production and conversion	40	34	41	36	43	69	49	66	39	67	50	68	56	61
[D] Cell cycle control, cell division, chromosome partitioning	35	36	41	37	43	37	33	40	24	33	23	45	30	38
[E] Amino acid transport and metabolism	112	106	159	137	160	192	152	129	110	136	116	179	139	152
[F] Nucleotide transport and metabolism	64	61	77	74	73	91	88	85	71	88	78	97	82	100
[G] Carbohydrate transport and metabolism	61	61	69	63	74	168	123	155	80	172	138	156	120	162
[H] Coenzyme transport and metabolism	51	49	54	49	64	91	73	80	52	72	64	98	78	78
[I] Lipid transport and metabolism	40	43	44	43	51	62	56	71	40	71	59	64	58	57
[J] Translation, ribosomal structure and biogenesis	180	175	188	180	190	193	191	185	162	193	166	198	186	191
[K] Transcription	93	84	89	87	115	133	128	129	93	150	132	153	100	151
[L] Replication, recombination and repair	110	86	97	86	115	110	100	105	57	92	95	119	96	125
[M] Cell wall/membrane/envelope biogenesis	84	77	73	74	84	110	92	105	81	98	75	102	93	94
[N] Cell motility	10	7	6	4	11	11	12	14	7	12	5	17	13	12

Table 3.2. Gene content profiles obtained for *Fructobacillus* spp. and *Leuconostoc* spp. (Continued)

	F. fructosus NRIC 1058^T	*F. durionis* DSM 19113^T	*F. ficulneus* JCM 12225^T	*F. pseudoficulneus* DSM 15468^T	*F. tropaeoli* F214-1^T	*L. mesenteroides* ATCC 8293^T	*L. carnosum* JB16	*L. citreum* KM20	*L. fallax* KCTC 3537^T	*L. gelidum* JB7	*L. inhae* KCTC 3774^T	*L. kimchii* IMSNU 11154^T	*L. lactis* KACC 91922	*L. pseudomesenteroides* 1159
[O] Posttranslational modification, protein turnover, chaperones	46	37	47	40	49	63	59	59	39	54	44	67	46	58
[P] Inorganic ion transport and metabolism	49	48	51	54	54	81	70	77	46	61	56	83	63	70
[Q] Secondary metabolites biosynthesis, transport and catabolism	10	7	12	9	12	18	10	13	10	11	12	11	15	15
[R] General function prediction only	67	55	78	67	85	99	83	87	64	89	77	103	79	95
[S] Function unknown	111	100	90	94	114	133	109	122	95	116	108	124	107	118
[T] Signal transduction mechanisms	31	27	36	29	36	60	49	55	46	48	44	60	51	58
[U] Intracellular trafficking, secretion, and vesicular transport	15	12	11	15	24	12	15	11	12	15	10	14	14	12
[V] Defense mechanisms	34	23	37	37	26	35	37	35	24	47	43	52	35	59
[X] Mobilome: prophages, transposons	44	12	26	9	33	27	21	42	18	12	51	43	38	58

To understand gene contents involved in metabolic/biosynthesis pathways in more detail, ortholog assignment and pathway mapping against the KEGG Pathway Database were performed using the KAAS system. The number of mapped genes was significantly less for *Fructobacillus* spp. as compared to *Leuconostoc* spp. (Table 3.3). Firstly, *Fructobacillus* spp. lack respiration genes. Whereas oxygen is known to enhance their growth [79], the strains have lost genes for the TCA cycle, and keep only one gene for ubiquinone and other terpenoid-quinone biosynthesis. Presumably they do not perform respiration and use oxygen only as an electron acceptor. This characteristic is not applicable to certain *Leuconostoc* species: *L. gelidum* subsp. *gasicomitatum* [168], formerly classified as *L. gasicomitatum* [169], has been reported to conduct respiration in the presence of heme and oxygen [170].

Secondly, *Fructobacillus* spp. lack pentose and glucuronate interconversions. They lost genes for pentose metabolism, unlike other obligately heterofermentative LAB that usually metabolize pentoses [171]. They do not metabolize mannose, galactose, starch, sucrose, amino sugars or nucleotide sugars, either [79,153]. Moreover, the species possess none or at most one enzyme gene for the phosphotransferase systems (PTS), significantly less than the number of respective genes in *Leuconostoc* spp. (13 ± 3.13, average ± SD). This validates the observation that *Leuconostoc* spp. metabolize various carbohydrates whereas *Fructobacillus* spp. do not [79] (Figure 3.4). However, the genome-based prediction does not always coincide with observed metabolism: *Fructobacillus* species do not metabolize ribose [79], against its metabolic prediction. The discrepancy is due to an absence of ATP-dependent ribose transporter. On the other hand, some *Leuconostoc* spp. have the transporter and metabolize ribose.

Thirdly, *Fructobacillus* spp. have more genes encoding phenylalanine, tyrosine and tryptophan biosynthesis compared to *Leuconostoc* spp., although this difference is statistically not significant ($p = 0.165$). The difference is mainly due to presence/absence of tryptophan metabolism, and the production of indole and chorismate. This is important to wine lactobacilli [172]. The reason of the sporadic conservation of indole biosynthesis in *Fructobacillus* remains unknown. In general, *Fructobacillus* spp. conserve relatively large number of genes involved in amino acid metabolism for their small genomes (Class E in Figure 3.3), which is well contrasted to *Lactobacillus delbrueckii* that lost large number of genes for amino acid biosynthesis during the adaptation to the protein-rich milk environment [155].

Table 3.3. Discriminative pathways between *Fructobacillus* spp. and *Leuconostoc* spp.

	***Fructobacillus* spp. Mean (SD)**	***Leuconostoc* spp. Mean (SD)**	***p***
Glycolysis (map00010)	12.2 (0.84)	19.5 (1.72)	0.001
TCA cycle (map00020)	0	4.2 (0.79)	
Pentose and glucuronate interconversions (map00040)	3.2 (1.64)	7.9 (2.80)	0.008
Fructose and mannose metabolism (map00051)	2.8 (0.84)	9.4 (2.12)	0.001
Galactose metabolism (map00052)	5.8 (0.84)	11.6 (2.72)	0.003
Ubiquinone and other terpenoid-quinone biosynthesis (map00130)	1 (0)	7.6 (0.97)	0.001
Oxidative phosphorylation (map00190)	9.2 (0.45)	12.7 (1.57)	0.001
Valine, leucine and isoleucine degradation (map00280)	2 (0)	4.4 (0.84)	0.001
Starch and sucrose metabolism (map00500)	6.4 (1.52)	12.9 (2.28)	0.001
Amino sugar and nucleotide sugar metabolism (map00520)	11.2 (0.45)	19.5 (2.17)	0.001
Pyruvate metabolism (map00620)	12 (1)	19.8 (1.99)	0.001
Carbon metabolism (map01200)	30.6 (3.21)	37.4 (3.20)	0.005
ABC transporters (map02010)	33.8 (3.11)	50.6 (8.34)	0.003
Phosphotransferase system (map02060)	1 (0)	13 (3.13)	0.03

Numbers shown in parenthesis correspond to the pathway map numbers in KEGG.

The values indicate means and standard deviations (in parenthesis) of numbers of genes used for the pathways.

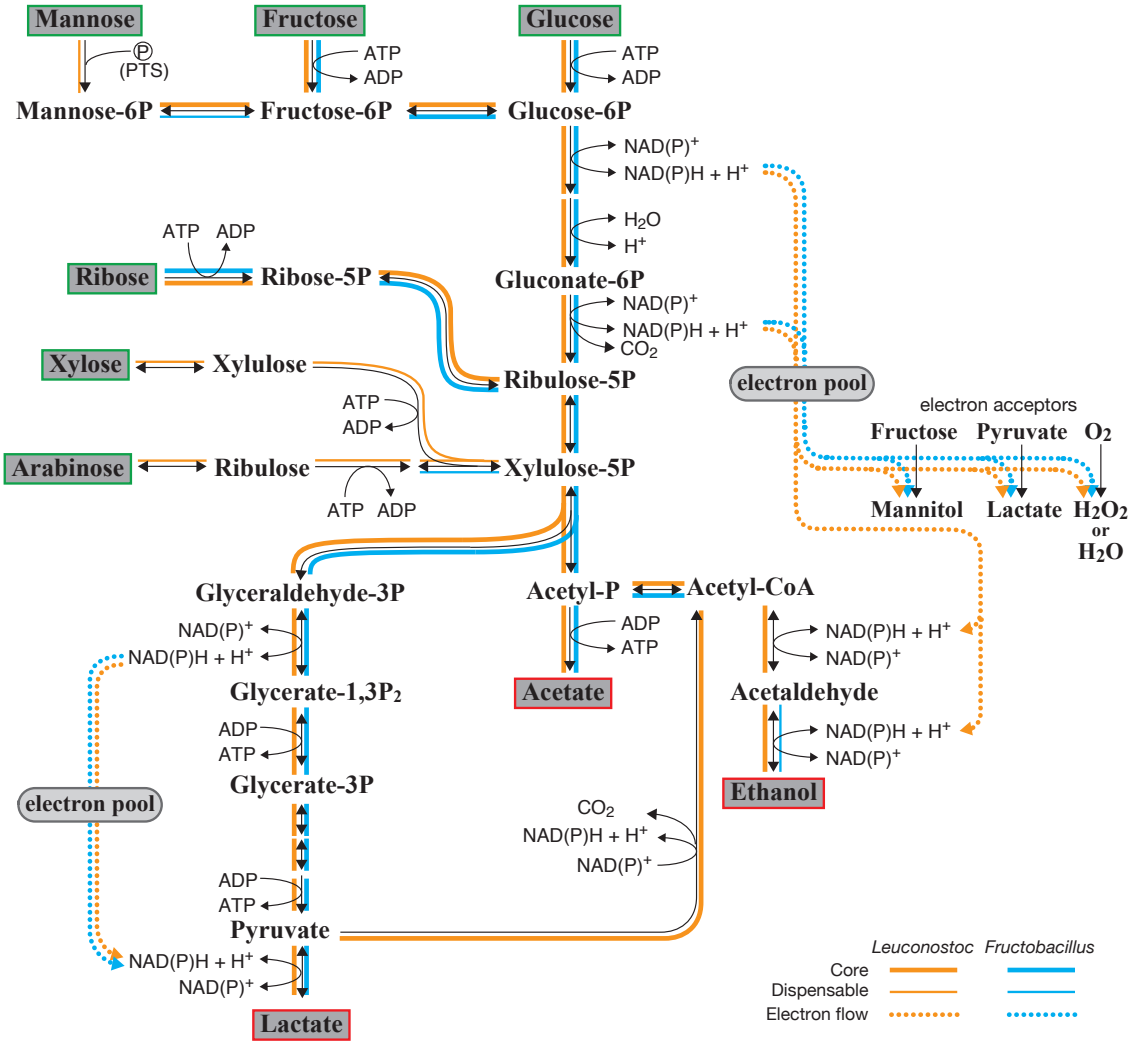


Figure 3.4. Predicted sugar metabolic pathways in *Fructobacillus* spp. and *Leuconostoc* spp. The bold lines represent core genes conserved in all the strains in each genus, while narrow lines represent dispensable genes that are absent in some strains. Dotted lines represent electron flow.

3.3.4. *Comparison of genus-specific genes*

To further investigate their differences, genes are defined as *Fructobacillus*-specific when they are conserved in four or more *Fructobacillus* species (out of five) and two or less in the nine *Leuconostoc* species. Likely, genes are *Leuconostoc*-specific when they are possessed by seven or more *Leuconostoc* species (out of nine) and zero or one in the five *Fructobacillus* species. According to this definition, 16 genes were identified as *Fructobacillus*-specific and 114 as *Leuconostoc*-specific (Table 3.4). These numbers are smaller than the numbers of fully conserved genes in each genus (24 for *Fructobacillus* and

159 for *Leuconostoc*, Figure 3.2B), because genus-specific genes was defined after mapping them to the KEGG Orthology (KO) database; genes without any KO entry were not taken into consideration in the analysis.

Table 3.4. Genus-specific genes for *Fructobacillus* and *Leuconostoc*.

***Fructobacillus*-specific genes**	***Fructobacillus* (n=5)**	***Leuconostoc* (n=9)**
alcohol dehydrogenase [EC: 1.1.1.1]	4	0
NAD(P)H dehydrogenase [EC: 1.6.5.2]	5	1
chloride peroxidase [EC: 1.11.1.10]	5	0
levansucrase [EC:2.4.1.10]	4	0
acylaminoacyl-peptidase [EC:3.4.19.1]	5	1
elaA; ElaA protein	5	0
MFS transporter, OPA family, glycerol-3-phosphate transporter	4	2
emrE; small multidrug resistance family protein	5	0
K06872 uncharacterized protein	4	0
K06994; putative drug exporter of the RND superfamily	4	2
K07025; putative hydrolase of the HAD superfamily	5	0
rsmD; 16S rRNA (guanine966-N2)-methyltransferase [EC:2.1.1.171]	5	0
cylA; multidrug/hemolysin transport system ATP-binding protein	5	0
ABC-2.CYL.P, cylB; multidrug/hemolysin transport system permease protein	5	2
adaA; AraC family transcriptional regulator, regulatory protein of adaptative response / methylphosphotriester-DNA alkyltransferase methyltransferase [EC:2.1.1.-]	4	0
pbuX; xanthine permease	5	1

***Leuconostoc*-specific genes (excerpt)**	***Fructobacillus* (n=5)**	***Leuconostoc* (n=9)**
pyruvate dehydrogenase E1 component subunit α & β [EC:1.2.4.1]	0	9, 8
dihydrolipoamide dehydrogenase [EC:1.8.1.4]	0	9
pyruvate dehydrogenase E2 component [EC:2.3.1.12]	0	9
aminopeptidase [EC:3.4.11.-]	0	7
dipeptidase D [EC:3.4.13.-]	0	7
Xaa-Pro dipeptidase [EC:3.4.13.9]	0	7
carboxypeptidase Taq [EC:3.4.17.19]	0	7
putative zinc metalloprotease [EC:3.4.24.-]	0	9
menaquinone-specific isochorismate synthase [EC:5.4.4.2]	0	9
celC; PTS system, cellobiose-specific IIA component [EC:2.7.1.69]	0	8
celA; PTS system, cellobiose-specific IIB component [EC:2.7.1.69]	0	8
celB; PTS system, cellobiose-specific IIC component	0	9
manX; PTS system, mannose-specific IIB component [EC:2.7.1.69]	0	9
manY; PTS system, mannose-specific IIC component	0	9
manZ; PTS system, mannose-specific IID component	0	9

ubiquinone/menaquinone biosynthesis methyltransferase [EC:2.1.1.163 2.1.1.201]	0	9
adhE; acetaldehyde dehydrogenase / alcohol dehydrogenase [EC:1.2.1.10 1.1.1.1]	0	9
tagG; teichoic acid transport system permease protein	0	8
tagH; teichoic acid transport system ATP-binding protein [EC:3.6.3.40]	0	8
cydC; ATP-binding cassette, subfamily C, bacterial CydC	0	8
cydD; ATP-binding cassette, subfamily C, bacterial CydD	0	7
tcyL; L-cystine transport system permease protein	1	7
tcyM; L-cystine transport system permease protein	1	7
tcyN; L-cystine transport system ATP-binding protein [EC:3.6.3.-]	1	7

Each number represents the number of species that possess the gene. An excerpt of 114 *Leuconostoc*-specific genes is shown.

Interestingly the *adh* gene coding alcohol dehydrogenase [EC:1.1.1.1] was characterized as *Fructobacillus*-specific whereas *adhE* gene coding bifunctional acetaldehyde/alcohol dehydrogenase [EC1.2.1.10, 1.1.1.1] was characterized as *Leuconostoc*-specific. There was no alternative acetaldehyde dehydrogenase gene in *Fructobacillus*. These results are consistent with the previous study reporting the lack of *adhE* gene and acetaldehyde dehydrogenase activity in *Fructobacillus* spp. [141] and their obligately heterofermentative nature with no ethanol production [79,152]. No production of ethanol is due to an absence of acetaldehyde dehydrogenase activity, but it conflicts with the NAD/NADH recycling. Therefore, there must be a different electron acceptor in glucose metabolism [141,150,152].

NAD(P)H dehydrogenase gene was found as *Fructobacillus*-specific. This is the only gene used for the quinone pool in *Fructobacillus* spp., suggesting that the gene does not contribute to respiration. Rather, it is used for oxidation of NAD(P)H under the presence of oxygen. This helps to keep the NAD(P)/NAD(P)H balance, since their sugar metabolism produces imbalance in NAD(P)/NAD(P)H cycling as described above. Although not *Fructobacillus*-specific, all strains of *Fructobacillus* possess NADH peroxidase, which also contributes to the NAD/NADH recycling as well as to cellular H_2O_2 detoxification. Indeed, *Fructobacillus* spp. can be easily differentiated from *Leuconostoc* spp. based on the reaction to oxygen [79]. In the validation study, *Fructobacillus* spp. grew well under aerobic conditions but poorly so under anaerobic conditions on GYP medium (Figure 3.5). Presence of oxygen had smaller impacts on growth of *Leuconostoc* spp., but they generated larger colonies under anaerobic conditions than under aerobic conditions.

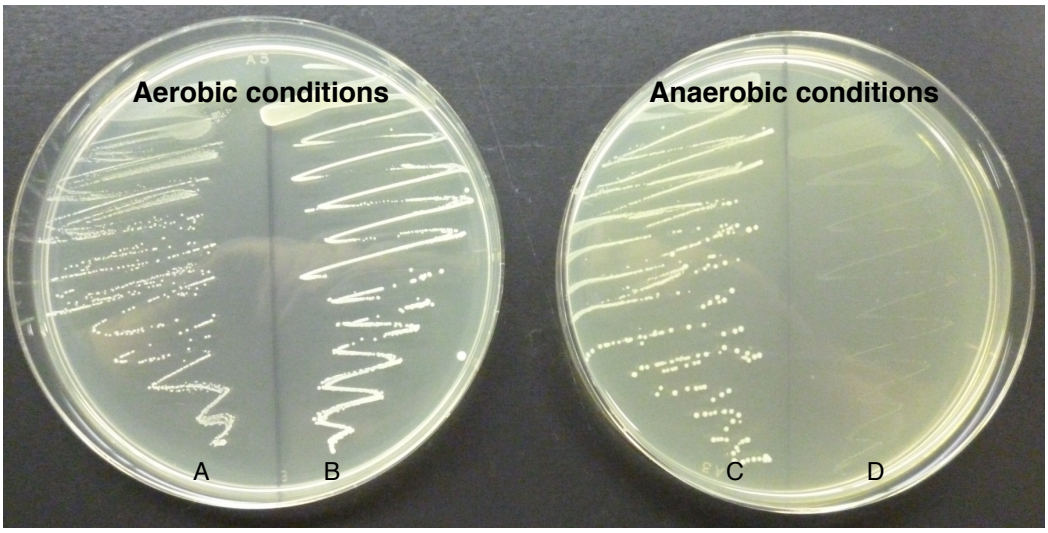


Figure 3.5. Growth of *L. mesenteroides* NRIC 1541^T and *F. fructosus* NRIC 1058^T on GYP agar medium under aerobic and anaerobic conditions after incubation for 2 days. A, C: *L. mesenteoides* NRIC 1541^T, **B, D**: *F. fructosus* NRIC 1058^T

Genes for subunits of the pyruvate dehydrogenase complex were undetected in the genomes of *Fructobacillus*, but were found as *Leuconostoc*-specific. *Fructobacillus* also lack TCA cycle genes. This suggests that, in *Fructobacillus*, pyruvate produced from the phosphoketolase pathway is not dispatched to the TCA cycle but metabolized to lactate by lactate dehydrogenase. The lack of pyruvate dehydrogenase complex was also reported in *Lactobacillus kunkeei* [173], which is also a member of FLAB found in fructose-rich environment [150].

The levansucrase gene was also characterized as *Fructobacillus*-specific. The enzyme has been known to work for production of oligosaccharides in LAB [174,175] and for biofilm production in other bacteria [176]. However, production of polysaccharides was unobserved in *Fructobacillus* spp. when cultured with sucrose. The reason for this discrepancy is yet unknown. Incompetence of sucrose metabolism, including no dextran production, in *Fructobacillus* spp. has been reported [79,153], and systems to metabolize sucrose, e.g. genes for sucrose specific PTS, sucrose phosphorylase and dextransucrase, were not detected in their genomes. On the other hand, *L. citreum* NRIC 1776^T and *L. mesenteroides* NRIC 1541^T produced polysaccharides, possibly dextran. Production of dextran from sucrose in the genus *Leuconostoc* is strain/species dependent [177], and dextransucrase gene was identified in six *Leuconostoc* genomes (out of nine) in this study. A number of genes coding peptidases and amino acids

transport/synthesis/metabolism were also found as *Leuconostoc*-specific genes, suggesting that *Leuconostoc* spp. can survive various environments with different amino acid compositions. Several PTS related genes and genes for teichoic acid transport were also characterized as *Leuconostoc*-specific. LAB cells usually contain two distinct types of teichoic acid, which are wall teichoic acid and lipoteichoic acid. The identified genes are involved in biosynthesis of wall teichoic acid in *Bacillus subtilis* [178]. Few studies have been reported for wall teichoic acid in *Leuconostoc* spp. and none in *Fructobacillus* spp.

3.3.5. *Phylogenetic analysis*

To confirm the phylogenetic relationship between *Fructobacillus* spp. and *Leuconostoc* spp., a phylogenetic tree was produced based on concatenated sequences of 233 orthologous genes which were conserved as a single copy within the tested strains. The tree showed a clear separation of the two genera (Figure 3.6), indicating that *Fructobacillus* spp. have distinct phylogenetic position from *Leuconostoc* spp. This agrees well with the previous reports using 16S rRNA gene or house-keeping genes [79,153].

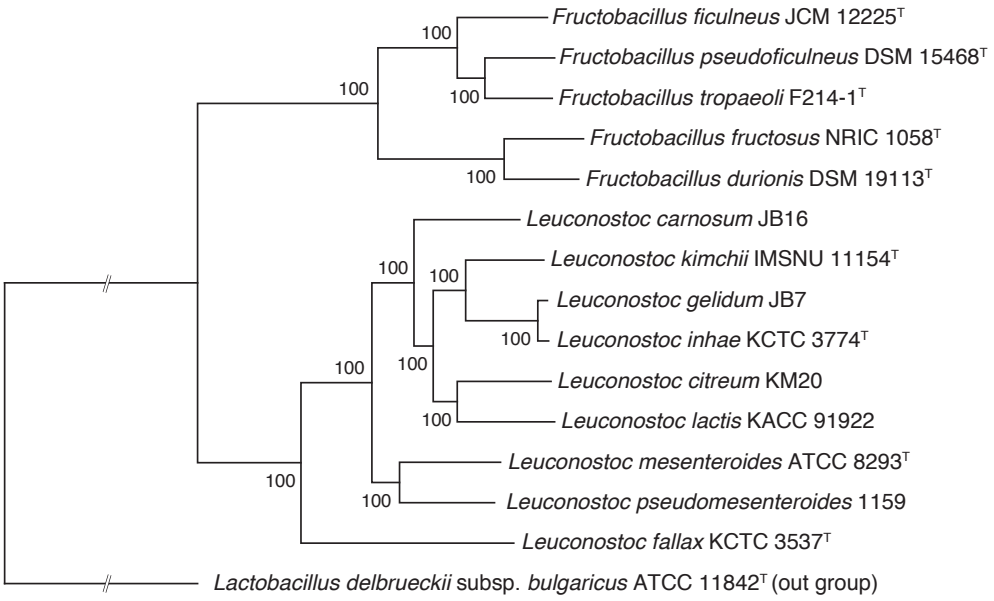


Figure 3.6. Phlylogenetic tree of *Fructobacillus* spp. and *Leuconostoc* spp. based on the multiple alignments of the 233 conserved genes. The partitioned maximum-likelihood tree was constructed using RAxML with the best-fit evolutionary model inferred by using FAsconCAT-G. The values on the branches are bootstrap support from 1,000 rapid bootstrapping replicates. *Lactobacillus delbrueckii* subsp. *bulgaricus* ATCC 11842^T was used as an outgroup.

3.3.6. *Selective advantage of Fructobacillus spp.*

Many of the plant-associated LAB strains are heterofermentative, able to utilize both pentose and hexose sugars. They tend to possess larger genomes to deal with the relatively large environmental fluctuation compared with the stable dairy environment. In particular, facultatively heterofermentative LAB such as *L. plantarum* and *L. paracasei* can be considered as "generalists" due to the versatility originating from their large genome sizes of about 3 Mbp. They can gain 2 ATPs both from homofermentation of a hexose and from heterofermentaion of a pentose. The high efficiency in energy production contributes to their predominance in many diverse environments. In contrast, obligate heterofermenters can obtain only one ATP per hexose, showing lower energy efficiency. In the unique metabolic pathway of *Fructobacillus* spp., although classified as obligately heterofermentative LAB, ethanol production in the conventional heterofermentative pathway is redirected to acetate production owing to the lack of the *adhE* gene and the ability to use fructose as an electron acceptor, which yields an extra ATP from a single-turnover of

hexose conversion (Figure 3.4). As shown in Figure 3.5, *Fructobacillus* can grow more vigorously than *Leuconostoc* under aerobic conditions by using oxygen as an electron acceptor. Together with energetical effectiveness of maintaining the small-size genomes, efficient acquisition of ATP might be a competitive advantage against other "generalist" microorganisms.

3.4. Conclusions

Genome-based analysis on conserved genes and metabolic characteristics clearly indicated the distinction between *Fructobacillus* spp. and *Leuconostoc* spp. *Fructobacillus* spp. possess smaller numbers of CDS in smaller genomes compared to *Leuconostoc* spp. This is mainly due to the absence of carbohydrate metabolic systems. Similar genomic characteristics have been reported for *L. kunkeei* [173], a member of FLAB found in fructose-rich environment. Since they are known as poor sugar fermenter in the group of LAB and always inhabit in fructose-rich niches, the characteristics could have resulted from an adaptation to their extreme environments. Niche-specific evolution, usually genome reduction, has been reported for dairy and vaginal LAB, and the present study reconfirms such niche-specific evolution in FLAB. These findings would be valuable to know a link of diverse physiological and biochemical characteristics in LAB and environmental factors in their habitats.

Chapter 4

4. Development of DFAST and DAGA: Web-based integrated genome annotation tools and resources

4.1. Introduction

As already described in Chapter 1, the genomic data for lactic acid bacteria (LAB) are now being generated all across the world and the number of genomes deposited at public sequence databases is increasing rapidly. DDBJ/ENA/GenBank are the core annotation databases of the International Nucleotide Sequence Database Collaboration (INSDC), collecting publicly available DNA information with metadata [52]. INSDC also collects raw sequences from the new-generation sequencing platforms into the Sequence Read Archive (SRA) [53]. These primary public databases constitute the basis for accessibility, reproducibility, and reusability of scientific data. However, since the quality assurance and correct assignment of taxonomy are the responsibility of data contributors, improving quality and taxonomic description has been an everlasting problem [40,59,178-180]. Low quality data not only lower the reliability of future analyses but also, in the worst case, lead to biologically incorrect conclusions. To avoid such problems, several tools and methods are available. QUAST [181] is a widely used assessment tool for genome assembly that reports statistical metrics such as N50 and detects misassemblies by using a reference genome. CheckM [92] estimates genome completeness and contamination by inspecting the presence/absence of marker genes specific to a given taxon. To confirm taxonomic affiliation of unidentified genome, Bull et al. proposed using 16S rRNA genes together with housekeeping genes [91]. Beaz-Hidalgo et al. recommended the use of average nucleotide identity (ANI) to verify the taxonomic position of the newly obtained genome [62]. ANI represents the mean of sequence identity of homologous regions in the alignment between a given pair of genomes, and an ANI value of 95–96% is widely accepted as the threshold for distinguishing species [31,182]. Examples of ANI values and the 16S rRNA gene sequences for curated genomes can be accessed at the EzGenome and EzTaxon databases [183]. According to the minimal standard recommended for describing new species of *Lactobacillus*, DNA-DNA hybridization (DDH) should be conducted if 16S rRNA sequence similarity to the closest known species is beyond 97% [184]. Recently, however, ANI has already been used as a substitute for DDH to describe novel species in *Lactobacillus* [36-38]. Furthermore, the use of genomic comparison methods

including ANI was also proposed at the workshop held at NCBI to find and correct misidentified genomes in the public databases [55].

Along this line of study, I have constructed a genome annotation pipeline called DDBJ Fast Annotation and Submission Tool (DFAST) and an associated repository, DFAST Archive of Genome Annotation (DAGA), both which are specialized for LAB. DAGA was developed to provide a reliable genome resource of LAB to the entire research community by assessing both quality and taxonomic affiliation of the genomic data. DAGA stores genome sequences reconstructed by *de novo* assembly of raw sequence data obtained from SRA as well as publicly available genomes obtained from DDBJ/ENA/GenBank, and all the genomes deposited in DAGA are consistently (re-)annotated with the newly developed annotation pipeline DFAST, thereby promoting accessibility and reusability of genome data. DFAST is based on an annotation pipeline Prokka [185] and a curated reference protein database tailored for LAB. DFAST is also equipped with quality and taxonomy assessment methods using CheckM and ANI. I also developed the user interface to provide metadata required for data submission to INSDC through the DDBJ Mass Submission System (MSS) [186] as well as to edit annotated features, and made it open to the public as a web service that realizes accurate and rapid genome analyses.

The initial version of DFAST and DAGA targeted LAB in the family *Lactobacillaceae*, which includes the genus *Lactobacillus*, the largest and diverse group comprising nearly 200 species and subspecies. *Lactobacillus* contains many species that have undergone reclassification as well as species difficult to distinguish by 16S rRNA gene sequences. The genus *Pediococcus* is another member of *Lactobacillaceae*, and is phylogenetically placed within the *Lactobacillus* cluster, near *L. plantarum* and *L. brevis* [70,74]. The term *Lactobacillus sensu lato* was also proposed to denote both of the two genera [75].The data stored in DAGA will be useful for all researchers who use LAB genomes, especially those focusing on inter- and intraspecific relations. In addition, as the showcases of data analyses benefiting from genomes deposited in DAGA, previously unreported intraspecific diversity in several *Lactobacillus* species and the niche-specific dissemination of genes among LAB strains are described. Both DFAST and DAGA are freely accessible at https://dfast.nig.ac.jp.

4.2. Methods

4.2.1. *Construction of the annotation pipeline*

The reference protein database for LAB was first constructed to provide consistent annotation to all genomes. A total 69 complete genomes of *Lactobacillus* and *Pediococcus*, publicly available as of

September 2015, were collected from the NCBI Assembly Database, and their protein sequences were extracted. In addition, 12 genomes were included to link with the *Lactobacillales*-specific Clusters of Orthologous Genes (LaCOGs) [64] and Microbial Genome Database (MBGD) [187]: *Aerococcus urinae* ACS-120-V-Col10a (GCA_000193205.1), *Carnobacterium* sp. 17-4 (GCA_000195575.1), *Enterococcus faecalis* V583 (GCA_000007785.1), *Lactococcus lactis* subsp. *cremoris* SK11 (GCA_000014545.1), *Lactococcus lactis* subsp. *lactis* Il1403 (GCA_000006865.1), *Leuconostoc mesenteroides* subsp. *mesenteroides* ATCC 8293 (GCA_000014445.1), *Melissococcus plutonius* ATCC 35311 (GCA_000270185.1), *Oenococcus oeni* PSU-1 (GCA_000014385.1), *Streptococcus pyogenes* M1 GAS (GCA_000006785.1), *Streptococcus thermophilus* LMD-9 (GCA_000014485.1), *Tetragenococcus halophilus* NBRC 12172 (GCA_000283615.1), and *Weissella koreensis* KACC 15510 (GCA_000219805.1). The identified 183,469 protein sequences were grouped into 28,002 orthologous clusters by using the GET_HOMOLOGUES software (version 1.3) with the default settings [161]. Briefly, candidates for orthologous genes were determined by bidirectional BLASTP alignments between each pair of the strains with an E-value threshold of 10e-5 and a minimum coverage threshold of 75%. Then, orthologous clusters were detected by the OrthoMCL algorithm. Among them, 11,993 were shared clusters containing two or more protein sequences, and the remaining 16,009 singletons were discarded. To infer the protein names and gene symbols, the shared clusters were mapped to the orthologous clusters of LaCOGs and MBGD. In total, 6,428 clusters were assigned to LaCOGs, of which 98.9% were consistently assigned to specific LaCOG clusters. Likewise, 1,601 clusters were assigned to MBGD, of which 94.4% were assigned consistently. To confirm the protein functions, public protein databases and the NCBI Conserved Domain Database [188] were searched manually. Protein names were determined following the NCBI guidelines for naming proteins (http://www.ncbi.nlm.nih.gov/genbank/genomesubmit_annotation/). The reference protein database was constructed from the 11,993 curated clusters. Figure 4.1 shows the schematic representation of the construction procedure.

The core annotation was based on the Prokka annotation software [185], performing prediction of tRNAs, rRNAs, CRISPRs, and protein-coding sequences as well as similarity searches against protein sequence databases and protein family profiles. The LAB reference database was used in our customized Prokka pipeline that can generate DDBJ-compliant submission files.

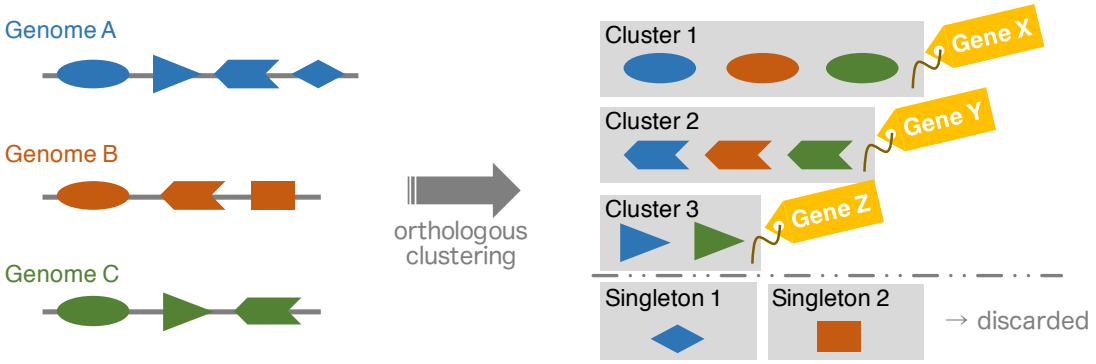


Figure 4.1. Schematic representation of the procedure for construction of the reference protein database. Protein sequences were grouped into orthologous clusters. Then each cluster was assigned a protein name. The reference protein database was constructed from the shared clusters.

4.2.2. *Data collection*

First, 743 publicly available genome sequences for *Lactobacillus* and *Pediococcus* were downloaded from the NCBI Assembly Database, which is a secondary database of DDBJ/ENA/Genbank that provides assembled sequences for each genome [58]. In addition, 678 raw sequence data (Illumina sequences with the paired-end method) were downloaded from SRA and assembled into contigs using the Platanus assembler (version 1.2.4) after preprocessing the reads using Platanus_trim (version 1.0.7) [98]. As Platanus was originally developed for heterozygous diploid genomes, the parameters "-d 0.3 -u 0.05" were specified to configure for bacterial haploid genomes. For each genome, *de novo* assembly was repeated five times by randomly sampling read sequences of different coverage, and the best result was chosen by the completeness calculated using CheckM and the average sequence length. All genomes were annotated with the newly developed DFAST pipeline. The taxonomic affiliations of the genomes were assessed by calculating ANI between 185 representative genomes whose taxonomic positions were confirmed (see Results and discussion). The workflow for data collection is shown in Figure 4.2.

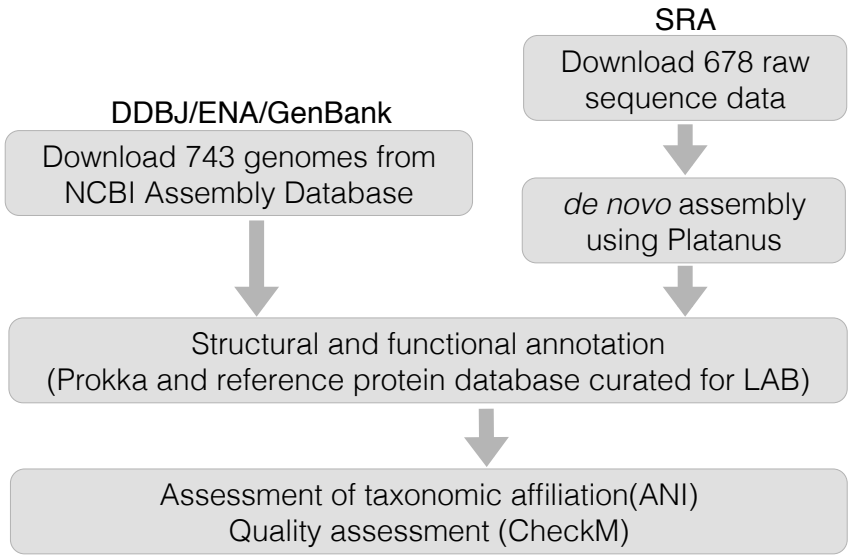


Figure 4.2. Data collection workflow for DAGA.

4.2.3. *Calculation of average nucleotide identity*

The pyani script (https://github.com/widdowquinn/pyani) was modified and used to calculate ANI between two genomes based on the method by Goris et al. [31]. In brief, one genome was cut into 1,020 nt fragments, which were searched against the other genome by using the BLASTN algorithm [189]. ANI was calculated as the mean identity of top-hit BLASTN matches for all fragments with a sequence identity of ≥30% and an overall aligned region of ≥70% of the fragment length. The trees in Figure 4.8 were constructed by the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) clustering method with a distance of (1 – ANI).

4.2.4. *Quality assessment of genomes*

CheckM (version 1.0.5) was used to calculate completeness and contamination of each genome [92]. CheckM inspected for the presence/absence of 409 and 664 single-copy gene markers specific for *Lactobacillus* and *Pediococcus*, respectively. Genome completeness and contamination were estimated by the number of distinct markers and their multiplicity in each genome, respectively.

4.2.5. *Phylogenetic analysis*

The phylogenetic tree shown in Figure 4.5 was constructed in the same manner as described in the Subsection 3.2.5 using 132 conserved single-copy genes among 185 representative strains and *Lactococcus lactis* as an outgroup.

4.2.6. *Gene transfer analysis*

The gene transfer among LAB strains was depicted as a network graph following the procedure below. First, protein coding sequences (CDS) in one genome were aligned to another using BLASTN algorithm. CDSs of transposase genes and conserved genes determined using 186 representative genomes were excluded prior to the alignment. Sequences aligned with the sequence identity ≥95% and overall aligned region of ≥95% were taken into consideration as candidate genes obtained by horizontal transfer. An edge was created between two genomes when the number of such candidates was more than or equal to 10 in a consecutive region of the genome. When two genes were within neighboring three loci in the genome, they were regarded as placed in a consecutive region. To exclude possible gene transfer from the common ancestors (vertical gene transfer), edges were eliminated when ANI values exceeds 77% between the two genomes. CytoScape (version 3.1.1) [190] was used to create the graph visualization.

4.2.7. *Implementation of the web service*

DFAST and DAGA were implemented in Python 2.7.11, PostgreSQL 8.4.20, and Nginx 1.8.0, and run on a Red Hat Enterprise Linux server (release 6.7). The job queuing system was developed using RabbitMQ 3.6.0.

4.3. Results and discussion

4.3.1. *Overview of the DAGA service*

The recent new generation sequencing technologies produced ever more genome sequences, making it important to assess their data quality and taxonomic positions. I developed an integrated genome archive specialized for LAB, namely DAGA (DFAST Archive of Genome Annotation), that stores quality-controlled and taxonomically confirmed bacterial genomes with consistent annotation. The first version of the datasets targeted the family *Lactobacillaceae* and contains 1,389 and 32 genomes for

Lactobacillus and *Pediococcus*, respectively. Among them, 743 are publicly available genome sequences deposited in DDBJ/ENA/GenBank; they were obtained from the NCBI Assembly Database. The remaining 678 genomes were assembled *de novo* from raw reads deposited in SRA. All genomes were annotated by the Prokka (ver. 1.11) pipeline with the custom reference database for LAB. As of January 2016, DAGA covers 180 species (including 19 subspecies) of the genus *Lactobacillus* and 11 species of the genus *Pediococcus*, which correspond to 91 % of the known species for both genera. DAGA utilizes accession numbers from the original source as the genome identifiers; data with "GCA" the genome identifiers are from the NCBI Assembly Database, and those with "DRR", "ERR", or "SRR" are from SRA. The completeness and contamination of genomes were assessed by examining the presence of specific gene markers with CheckM, which could successfully identify genomes of incorrect size as compared with typical LAB strains without using any other selection method. The genome completeness values partly depend on the sequencing platform, as some sequencers are more prone to insertion/deletion errors than others [17], making gene calling difficult due to the frameshifts. The taxonomic affiliation of each genome was verified by calculating ANI. I could find taxonomically mislabeled genomes in the public database even for type strains, which will be discussed later. Of note, the genus *Sharpea* was not included even though it is still classified in the family *Lactobacillaceae*. *Sharpea azabuensis*, the only member of this genus, was initially described as a species related to *Lactobacillus catenaformis*, but *L. catenaformis* was later reclassified as *Eggerthia catenaformis*, and it is no longer a member of *Lactobacillaceae* [191,192]. As the number of available genomes is increasing rapidly, I plan to update the database regularly and to expand the scope of the database to other taxonomic groups.

Figure 4.3 shows screenshots of DAGA. Users can query genomes of interest from the search form in the upper part, or select taxonomic name. A keyword search is available too. The genome quality is rated in 5 grades, allowing users to easily select reliable genomes for comparative analysis. The definition of the rating scale and the number of genomes in each grade are shown in Table 4.1. DAGA also provides genome statistics: the number of coding sequences, estimated genome size, and external links to related databases. Annotation results can be downloaded in either GenBank or FASTA format files. DAGA is freely accessible at https://dfast.nig.ac.jp.

Likewise, NCBI Reference Sequence (RefSeq) and the Pathosystems Resource Integration Center (PATRIC) provide consistently annotated genome collections [193,194]. They collect genome sequences from DDBJ/ENA/GenBank and re-annotate them using NCBI Prokaryotic Genomes Annotation Pipeline (and Rapid Annotation using Subsystem Technology (RAST), respectively. As far as we know, there is no database that collects genomic data from both DDBJ/ENA/GenBank and SRA. Since SRA stores raw

sequence data, it is difficult for users incapable of bioinformatics analysis to exploit the data. Also, some data are only available in SRA, for example the sole reliable genome for *L. amylotrophicus* was obtained from SRA (ERR387486). DAGA facilitates the reuse of the valuable data mined in SRA.

Table 4.1A. Number of genomes deposited in DAGA.

Data Source	Quality Rating					Total
	1	2	3	4	5	
DDBJ/ENA/GenBank	17	11	59	558	98	743
SRA	30	27	4	617	0	678
Total	47	38	63	1,175	98	1,421

Table 4.1B. Definition of the quality rating grades.

Quality Rating	Definition
5	High Quality Complete Genomes with completeness ≥ 95% and contamination ≤ 5%
4	High Quality Draft Genomes with completeness ≥ 95% and contamination ≤ 5%
3	Low Quality Genomes with completeness ≥ 80% and contamination ≤ 10%
2	Disqualified Genomes with completeness < 80% or contamination > 10%
1	Taxonomically mislabeled or misidentified Genomes

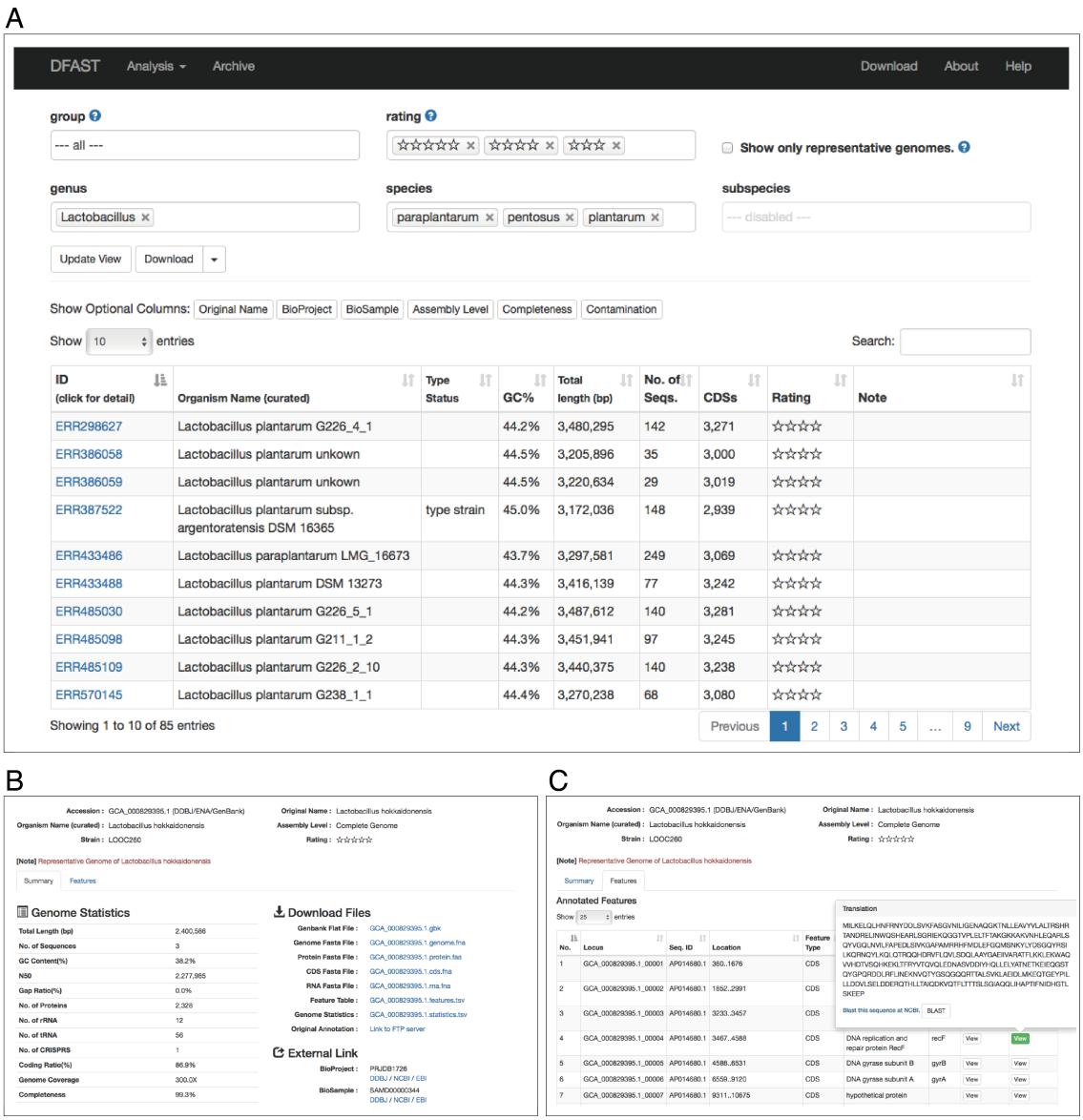


Figure 4.3. Screenshots of DAGA. A: Main page of DAGA, listing genomes in the database. Users can query genomes from the search form. **B**: Detail page of each genome, showing statistics and external links. Data files are downloadable in several formats. **C**: Detail page of annotated features. Links to the BLAST web service at NCBI are available.

4.3.2. *Selection of a representative genome for each LAB species*

To verify the taxonomic relationship of LAB, pairwise ANI values were calculated among 191 strains representing each species (or subspecies). Priority was given to the type strains in the data selection, and when multiple genomes were available, the one with the highest completeness and the longest average

sequence length was chosen. Figure 4.4 shows the results of ANI calculation (also see the website https://dfast.nig.ac.jp/download). In most cases, the interspecific ANI values were below 95%, the threshold to differentiate species. Six strains listed in Table 4.2 showed anomalously high ANI values between different species (red circles in the Figure 4.4B), indicating the incongruence of their taxonomic positions, which will be discussed in the following subsection.

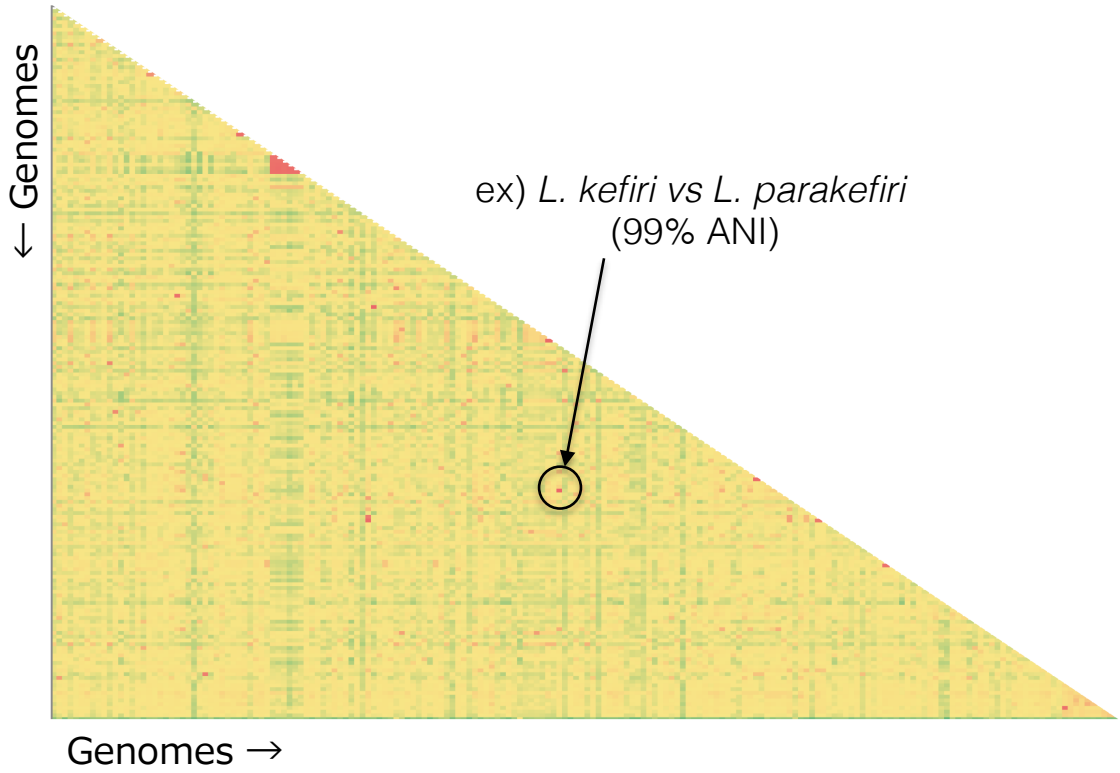


Figure 4.4A. Plot for All-vs-All ANI calculations among 191 LAB species. Each dot represents the ANI value of given pair of genomes. Red represents higher values and green represents lower values. An example of anomalous values is shown for the one between *L. kefiri* and *L. parakefiri*.

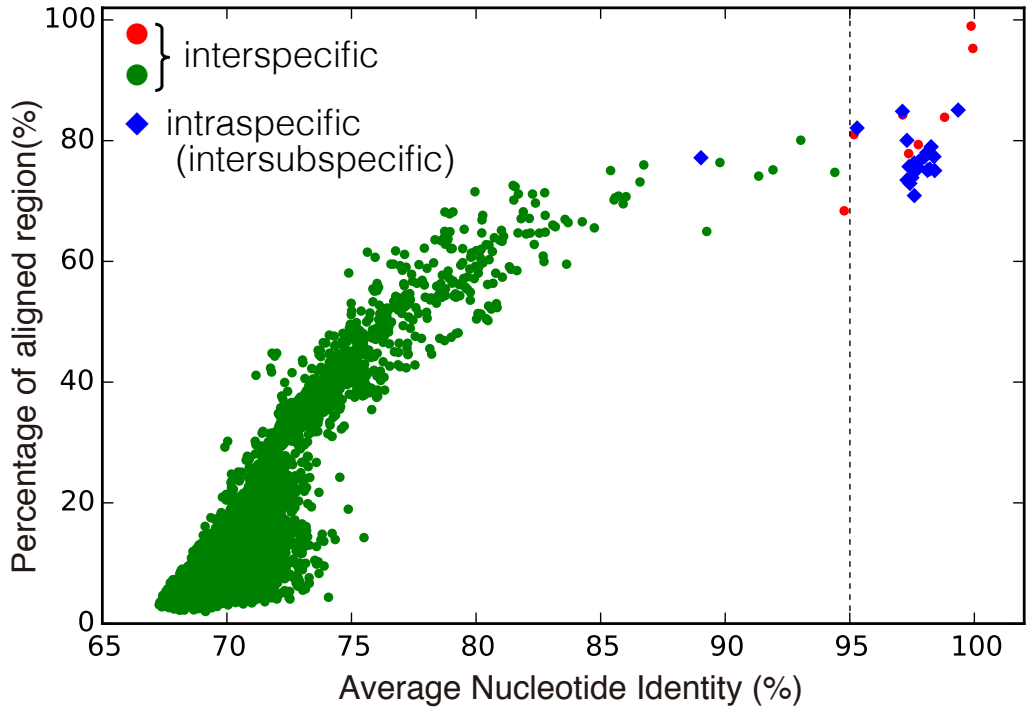


Figure 4.4B. Distribution of All-vs-All ANI values. The green and red circle both represent interspecific ANI values. The red ones represent anomalous values from six strains listed in the Table 4.2. The blue diamonds represent intraspecific ANI values (between different subspecies in a species).

Table 4.2. Strains with anomalously high ANI values

Data Source*	Organism Name	Description
GCA_000159175.1	*Lactobacillus brevis* subsp. *gravesensis* ATCC 27305#	Shows 97.3% ANI value against *L. hilgardii*.
ERR387492	*Lactobacillus fornicalis* JCM 12512^T	Shares 98.7% ANI with *L. plantarum* subsp. *plantarum*.
GCA_001436985.1	*Lactobacillus homohiochii* DSM 20571^T	Shares 99.9% with *L. fructivorans*.
GCA_001434215.1	*Lactobacillus parakefiri* DSM 10551^T	Shares 99.9% ANI with *L. kefiri*. Possibly, contaminated with *L. kefiri* (contamination value 98%).
SRR1561417	*Pediococcus lolii* DSM 19927^T	Shares 97.1% with *P. acidilactici*.
GCA_001437265.1	*Pediococcus parvulus* DSM 203321^T	Shares 92.5% with *P. acidilactici*. Possibly, contaminated with *P. acidilactici* (contamination value 98.9%).

non-type strain

Figure 4.4B also indicates prominent discriminatory power of ANI to distinguish two LAB species. Only 0.4% of the comparisons fell within the "twilight zone" of 85-95% ANI values. Even between

hard-to-distinguish taxonomic groups such as *L. casei* and *L. plantarum*, ANI values were below 85%, much less than the threshold of 95%. In addition, ANI does not require gene calling and is applicable to draft genomes. It is especially valuable in the case of conducting *de novo* assembly from short reads because bacterial genomes normally encode multiple rRNA operons, which makes it difficult to reconstruct rRNA sequences. Besides, ANI does not require a laboratory assay and is computationally reproducible. For these reasons, I emphasize the benefit of ANI to validate taxonomic status for genomes deposited in DAGA.

After excluding these six strains in Table 4.2, 185 representative genomes were determined whose interspecific pairwise ANI values were well below 95%. One exception was *L. zeae* DSM 20178^{T} and *L. casei* ATCC 393^{T}, which had an ANI of 94.4%. After a long period of controversy, *L. zeae* is now considered to be in the same taxon as *L. casei* [195]. As shown later in the subsection 4.3.6, ANI values between species were always less than 95% in our analysis, but the reverse was not always true; In some species, intraspecific ANI values can be lower than 95%. Thus, it would be quite pertinent to consider the two strains to belong to the same species. However, the name of *L. zeae* still has not been validly rejected in the current nomenclature, therefore, it was selected in the database with its original name. It should also be noted that the publicly available genome for *L. amylotrophicus* (GCA_001434555.1), which exhibited an ANI of 99.9% with *L. amylophilus*, did not serve as the representative genome. Instead, the one from SRA (ERR387486) was used as the representative of *L. amylotrophicus*. In a recent study, it was postulated that *L. amylotrophicus* was a later synonym of *L. amylophilus* [75]. This result re-justified the taxonomic classification of *L. amylotrophicus*.

The validity of the 185 representative genomes was also confirmed by comparing the reconstructed 16S rRNA gene sequences with those deposited in public databases. When not available, housekeeping genes like *pheS* or *rpoA* were used. In addition, a phylogenetic tree was constructed using 132 conserved single-copy genes among them to verify their taxonomic positions (Figure 4.5). Figure 4.6 shows the statistics of 185 representative genomes. The selection of representative genomes was implemented as a procedure in our system to serve as a tool for taxonomic study where comparison with type strains is critical.

Figure 4.5A. Phylogenetic tree of 185 representative LAB genomes. The maximum likelihood method was used to construct the tree based on the concatenation of multiple alignments of the 132 conserved genes. *Lactococcus lactis* was used as an out group. The tree with bootstrap values is available at the download site of the web service (https://dfast.nig.ac.jp/download/).

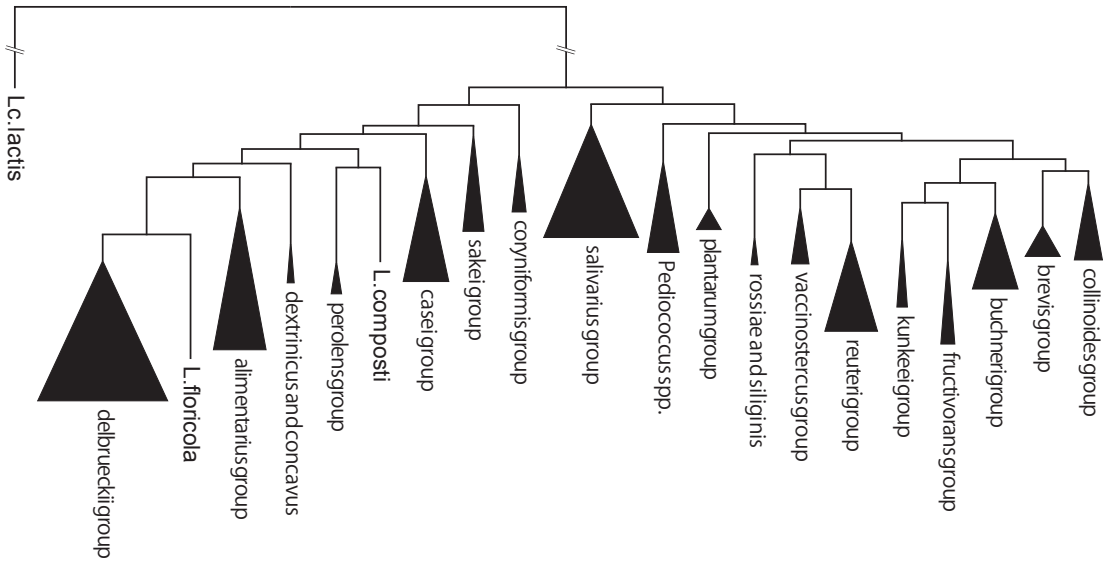


Figure 4.5B. Phylogenetic tree of 185 representative LAB groups. Triangles represent subtrees for taxonomic groups according to Felis, et al [196].

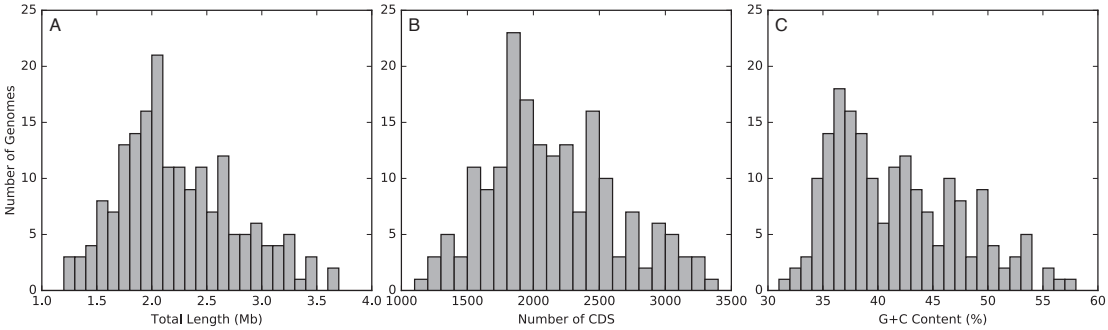


Figure 4.6. Distributions of genome statiscitcs for 185 representative genomes. A) estimated genome sizes, **B**) numbers of CDSs, **C**) GC contents.

4.3.3. *Taxonomic status of the six strains with anomalous ANI values*

The six strains listed in Table 4.2 were excluded from the representative genomes, namely, *Pediococcus lolii* DSM 19927^T (GCA_001437115.1), *Pediococcus parvulus* DSM 203321^T (GCA_001437265.1),

Lactobacillus brevis subsp. *gravesensis* ATCC 27305 (GCA_000159175.1), *Lactobacillus fornicalis* JCM 12512^T (ERR387492), *Lactobacillus homohiochii* DSM 20571^T (GCA_001436985.1), and *Lactobacillus parakefiri* DSM 10551^T (GCA_001434215.1). *P. lolii* was presumably misclassification of sequenced strains. A previous study reported that the type strains of *P. lolii* deposited in DSMZ and JCM were strains of *Pediococcus acidilactici* [197]. This analysis showed that not only *P. lolii* DSM 19927^T but also strain NGRI 0510Q^T (GCA_000319265.1), an original type strain of *P. lolii*, shared an ANI of 97% with *P. acidilactici*.

L. brevis subsp. *gravesensis* was first described over 60 years ago, but it was not mentioned in the Approved Lists of Bacterial Names published in 1980 [198]. This species is displayed as *Lactobacillus* sp. and *Lactobacillus hilgardii* in JCM and the EzGenome database, respectively [199,183]. The type strains of *L. homohiochii* and *L. fornicalis* deposited in culture collections were reported to misrepresent the originally described strains [200] (http://www.bacterio.net/lactobacillus.html#fornicalis). Their original strains are no longer available, and designation of a neotype seems appropriate.

The genome of *L. parakefiri* DSM 10551^T (GCA_001434215.1) exhibited an extremely high contamination value (98%), indicating the mixture of different strains. Indeed, two *pheS* genes were found in the genome, each matching the deposited *pheS* gene sequences of *L. kefiri* and *L. parakefiri*. In recent two studies, different statements were made for *L. parakefiri*; Zheng et al. argued that *L. parakefiri* was a later heterotypic synonym of *L. kefiri* [75], whereas Sun et al argued that it had the largest genome in the genus *Lactobacillus* [89]. But both studies seem to fail to describe the actual situation. Likewise, the genome of *P. parvulus* DSM 20332^T seemed to be contaminated with another strain of *P. acidilactici*.

4.3.4. *Detection of mislabeled genomes*

The taxonomic affiliation for all genomes in DAGA was validated by conducting ANI calculations against the representative genomes. Through this process, I corrected the species name for 77 mislabeled genomes, and inferred names for 55 unidentified genomes that were deposited as *Lactobacillus* sp. Table 4.3 shows examples of mislabeled genomes deposited in DDBJ/ENA/GenBank. Genomes with ambiguous taxonomic position were marked as Rating 1 (see Table 4.1B).

Table 4.3. Examples of mislabeled genomes deposited in DDBJ/ENA/GenBank.

Data Source*	Organism Name	Description
GCA_000159195.1	*Lactobacillus buchneri* ATCC 11577	Shares 99.1% ANI with *L. hilgardii*.
GCA_001434555.1	*Lactobacillus amylotrophicus* DSM 20534^T	Shares 100% ANI with *L. amylophilus*. Possibly, replaced by the strain of *L. amylophilus*.
GCA_001314245.1	*Lactobacillus gallinarum* HFD4	Shares 96.7% ANI with *L. helveticus*.
GCA_001273585.1	*Lactobacillus plantarum* SNU.Lp177	Shares 98.9% ANI with *L. plantarum* subsp. *argentoratensis* and 95.6% with subsp. *plantarum*.
GCA_001068345.1	*Lactobacillus johnsonii* 987_LJOH	Shares 93.4% ANI with *L. gasseri*.
GCA_001066235.1	*Lactobacillus johnsonii* 770_LJOH	Shares 100% ANI with *L. gasseri*.
GCA_001064985.1	*Lactobacillus helveticus* 459_LHEL	Shares 96.8% ANI with *L. gasseri*.
GCA_001063065.1	*Lactobacillus kefiranofaciens* 249_LKEF	Shares 100% ANI with *L. gasseri*.
GCA_001063045.1	*Lactobacillus crispatus* 240_LCRI	Shares 100% ANI with *L. gasseri*.
GCA_000469115.1	*Lactobacillus plantarum* AY01	Shares 99.6% ANI with *L. paraplantarum*.
GCA_000463075.2	*Lactobacillus plantarum* EGD-AQ4	Shares 92.8% ANI against *L. pentosus*.
GCA_000191545.1	*Lactobacillus acidophilus* 30SC	Shares 100% ANI against *L. amylovorus*.
GCA_000159195.1	*Lactobacillus buchneri* ATCC 11577	Shows 99.1% ANI against *L. hilgardii*.

Only genomes obtained from DDBJ/ENA/GenBank are listed. Genomes for *L. casei*/*paracasei* complex and unidentified genomes (*Lactobacillus* sp.) are not shown.

Twenty-eight out of 32 "*L. casei*" genomes were in fact *L. paracasei*, as previously postulated in the literature [81] and indicated by the fact that they shared an ANI of over 98% with *L. paracasei* ATCC 25302^T and an ANI of less than 85% with *L. casei* ATCC 393^T. Among the remaining four "*L. casei*" genomes, two were those of type strains, one was low quality with 22% ambiguous bases (N), and the last was recently published *L. casei* N87 (GCA_001013375.1). The last strain shared 96.8% ANI with *L. zeae* DSM 20178^T and 94.3% ANI with *L. casei* ATCC 393 ^{T}, indicating a genuine *casei* strain.

In the *L. plantarum* group, the members of which are notoriously difficult to identify with 16S rRNA sequence similarity, three "*L. plantarum*" genomes were reassigned organism names inferred from ANI results. The strains SNU.Lp177 (GCA_001273585.1), EGD-AQ4 (GCA_000463075.2), and AY01 (GCA_000469115.1) were *L. plantarum* subsp. *argentoratensis*, *L. pentosus*, and *L. paraplantarum*, respectively.

4.3.5. *DFAST on line annotation server*

I developed web user interfaces for the annotation pipeline used in this study, and released as an on-line annotation platform called the DDBJ Fast Annotation and Submission Tool (DFAST). Users can annotate their own genome sequences by uploading a FASTA formatted file via a submission form, and can perform quality and taxonomic assessment using CheckM and ANI calculation as well. A simple annotation editor is also available, allowing users to modify gene product names or gene symbols. Results can be downloaded in several different formats including GenBank flat file, Multi-FASTA, or tab-separated tables. In addition, users can manage metadata and create submission files for DDBJ Mass Submission System. Figure 4.7 shows representative screenshots of DFAST. DFAST is developed so that all the procedure required for submission can be done seamlessly on-line, thus it can be used as an on-line workspace to prepare submission files to DDBJ, which will be especially useful for users not familiar with bioinformatics skills.

In comparison with other annotation tools such as RAST or the Microbial Genome Annotation Pipeline (MiGAP) [101], the advantage of DFAST is the ability to generate ready-to-submit level annotation files. RAST can perform detailed functional annotation based on the platform known as SEED. However, if users want to submit an annotated genome to INSDC, they need to convert annotation results into acceptable formats. Although MiGAP partly supports the DDBJ-acceptable formats, users are required to provide metadata and to curate annotated protein names before submission. As the curated reference database constructed in this study followed the protein naming guidelines by NCBI, only minimal manual curation, if any, is required before submitting genomes to DDBJ. Short running time is another advantage of DFAST. It takes about 5 minutes to annotate a typical size bacterial genome, while RAST and MiGAP take several hours. In addition, DFAST provides quality and taxonomy assessment tools, which prevent users from submitting low-quality or mislabeled genomes to INSDC. I have already used DFAST to annotate and submit genomes of 5 LAB strains, including two candidates for new species (manuscript in preparation). On average, 90.3% of protein coding sequences were annotated based on similarity search results against the reference protein database in this study. Currently, DFAST is based on the simple annotation pipeline, Prokka, and thus does not provide functions to annotate frameshifted genes or pseudogenes, which will be a future issue of DFAST. Another future task is an update of the reference database, which is currently constructed mainly for *Lactobacillus* and *Pediococcus*, and does not fully support other genera such as *Lactococcus* or *Leuconostoc*. In addition, I have a plan to extend DFAST to organisms other than LAB.

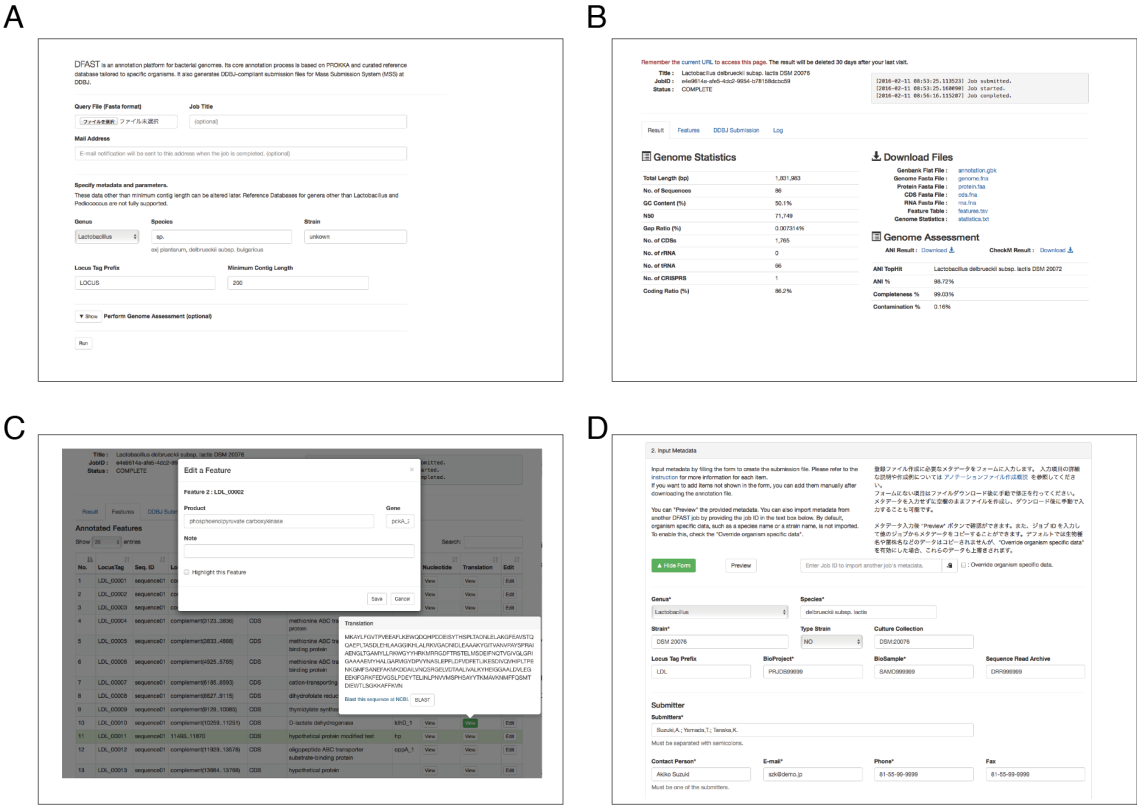


Figure 4.7. Screenshots of DFAST. A: Submission form of DFAST. Users can annotate their own genome by uploading Fasta file. **B**: Result of DFAST. Data files are downloadable in several formats. **C**: Detail page of annotated features. Links to the BLAST web service at NCBI are available. **D**: Submission files for DDBJ Mass Submission System can be generated by providing metadata.

4.3.6. *Intraspecific diversity of LAB revealed by ANI*

To further investigate genomic diversity of LAB, I conducted all-against-all ANI comparison between 1,336 genomes deposited in DDBJ/ENA/GenBank (N=1,336×1,335/2=891,780). Low-quality genomes and genomes with ambiguous taxonomy were excluded. Figure 4.8 shows the distribution of ANI values. All interspecific ANI values (N=862,221) were less than the species-delineation cutoff value of 95%, while 1,670 out of 29,559 intraspecific ANI values were also less than 95%. Such exceptions included *L. kunkeei*, *L. gasseri*, *L. jensenii* and *L. vaginalis*, suggesting the high intraspecific diversity. *L. gasseri* and *L. jensenii* were each separated into two previously unreported subgroups. Similarly, *L. vaginalis* could be separated into at least two intraspecific subgroups (Figure 4.9A–C). The ANI values between the subgroups were 93% and 88% for *L. gasseri* and *L. jensenii*, respectively, while ANI values within the same subgroups were over 98% in both species. The intraspecific separation was also supported by the

multiple alignments of *pheS* and *rpoA* gene sequences (data not shown).

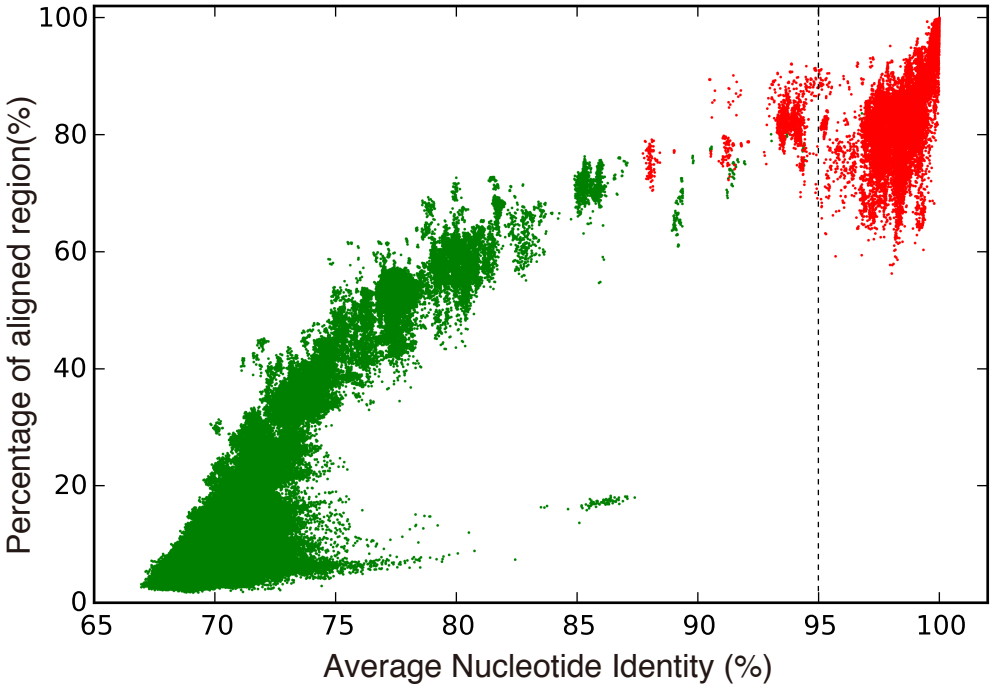


Figure 4.8. The green and red dots represent interspecific and intraspecific ANI values, respectively. Species whose intraspecific ANI values are less than the species-delineation cutoff of 95% include *L. kunkeei*, *L. gasseri*, *L. jensenii*, and *L. vaginalis*.

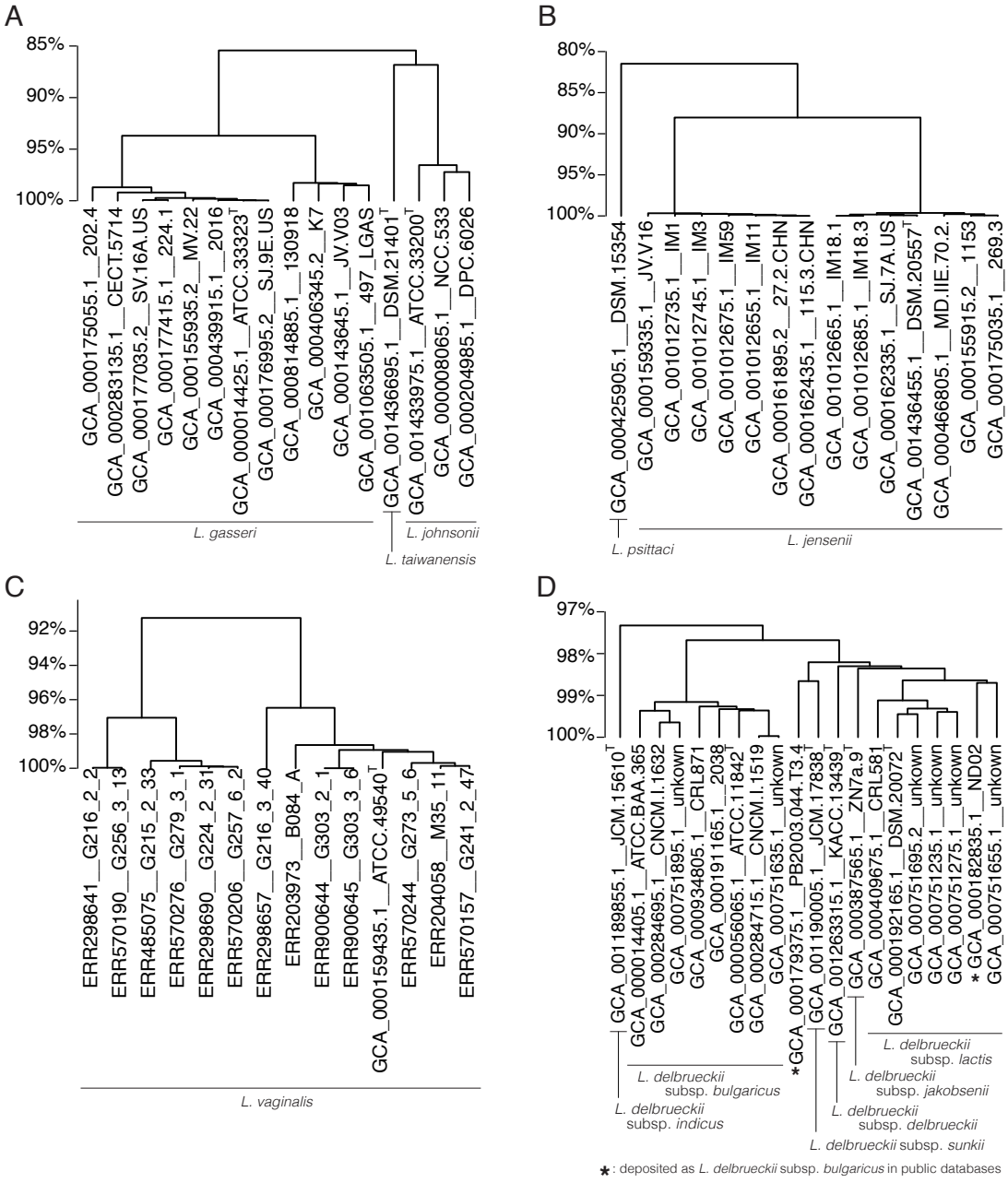


Figure 4.9. Hierarchical clustering results by using (1 – ANI) as the genome distance. A: *Lactobacillus gasseri,* **B**: *Lactobacillus jensenii*, **C**: *L. vaginalis*, **D**: *L. delbrueckii*. Each label represents the accession of the data source and the strain name.

The three species mentioned above (Figure 4.9A–C), in particular *L. jensenii*, are common habitants in human vagina. LAB are predominant organisms in human vaginal microbiota and the composition of microbial community is reported to depend on ethnic groups [201,202]. It is tempting to speculate on the

biogeographical and ethnic factors against their genetic variation. The intraspecific diversity of these species was well below the species-level threshold and might correspond to subspecies-level differentiation. It must be noted, however, that the analysis conducted here was based on genomic information only. Therefore, further analysis including polyphasic characterization is required to establish their valid classifications.

The ANI values between different subspecies differed widely. For example, *L. plantarum* and *L. aviarius* showed 95% and 90% intersubspecific ANI values, respectively, equal to or less than the threshold to distinguish species. By contrast, *L. paracasei* and *L. delbrueckii* are rather homogeneous with about 98% ANI values between subspecies. We could not find a unified demarcation line that discriminate subspecies. However, in spite of the high ANI values, hierarchical clustering based on the ANI values could separate five subspecies of *L. delbrueckii* (Figure 4.9D). The tree topology was roughly consistent with the ones from multi locus sequence analyses [203,204]. This implies the reliability of ANI in evaluating subtle variation within species. For other subspecies group, we could not perform sufficient analyses because of the limited number of genomes.

4.3.7. *Gene transfer among LAB*

The gene transfer among 606 LAB genomes in DAGA was depicted as a network graph shown in Figure 4.10, in which each node represents a genome and edges represent gene transfer between them. In this analysis, nodes were linked only when the number of genes possibly acquired via horizontal transfer was 10 or more in a consecutive region. In addition, the nucleotide identity threshold was set to 95% and gene transfer between closely related species was excluded by calculating ANI. The ANI cutoff value was initially set at 85% since 99.6% of the interspecific ANI calculations resulted in a lower value than this threshold (Figure 4.4B). However, when multiple genomes of isolates from the same species are available, too many edges were created, making it difficult to obtain informative meaning. For example, the number of edges linked between *L. paracasei* (79 genomes) and *L. rhamnosus* (57 genomes) was 127, which accounted for 27% of all the edges. So the threshold was empirically set at 77% to exclude edges between the two species. Therefore, this figure does not capture the whole perspective of gene transfer, but rather recent and relatively large-scale acquisition like genomic islands or plasmids between somewhat distantly related species. Nevertheless, I could identify remarkable dissemination of genes associated with anti-stress system among the strains from similar environmental conditions.

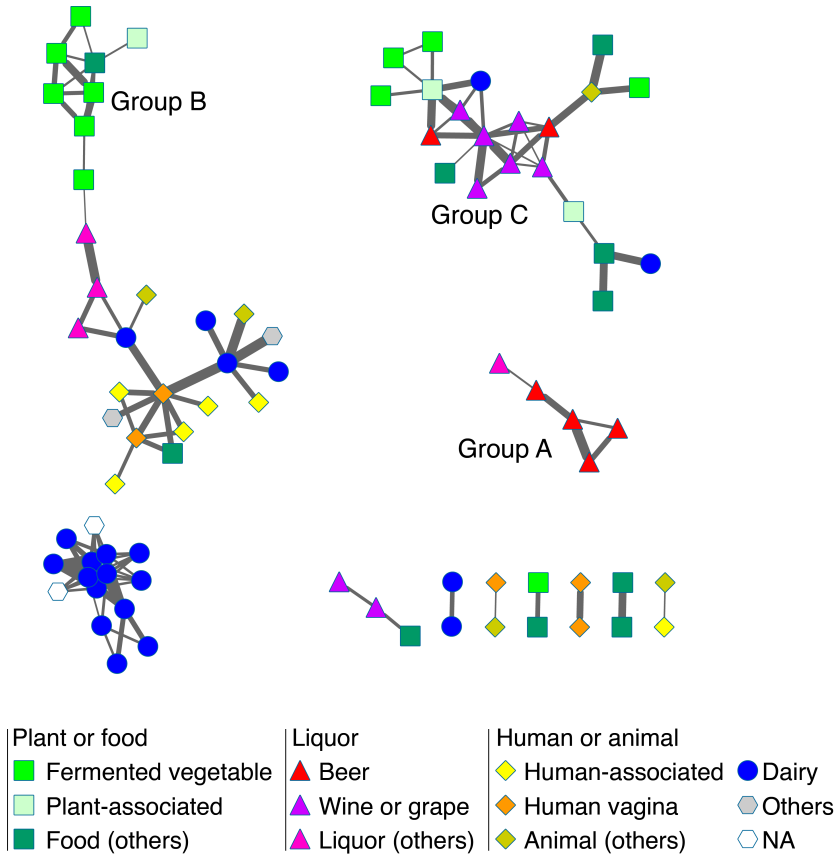


Figure 4.10. Gene transfer network among LAB strains. Each node stands for a genome with its isolation source represented by the shape and color. Edges represent gene transfer between genomes. The line width corresponds to the number of candidate genes obtained by horizontal transfer (see methods).

Group A in Figure 4.10 includes beer spoilage LAB sharing hop resistance genes. The horizontal acquisition of hop resistance genes have been reported in several studies and they are often mediated by plasmids or transposons [205]. The group A strains share the *horA* gene involved in hop resistance and the *gtf* genes responsible for exopolysaccharide synthesis and beer spoilage, which are localized on a plasmid of well-characterized beer contaminant strain, *L. brevis* BSO 464 [206]. To elucidate whether the gene transfer was limited within the family *Lactobacillaceae*, the *horA* genes were searched against NCBI non-redundant nucleotide/protein databases with excluding *Lactobacillaceae*. As a result, *horA* gene homologues were found with over 99% nucleotide identity in four Gram-positive bacterial isolates from *Staphylococcus*, *Bacillus*, and *Paenibacillus*. They were all isolated from spoiled beer and were reported by the same research group [212]. In general, isolates of these genera are not regarded as beer-spoilage organisms. Presumably, the *horA* genes were transferred from *Lactobacillus* or *Pediococcus*. Group B

strains are associated with traditional fermented foods, such as Kimchi in Korea and Suguki in Japan, and harbor genes for thioredoxin and thioredoxin reductase in common. The thioredoxin systems are shared not only within lactobacilli but within *Leuconostoc* spp. isolated from Kimchi, and they are also reported to be encoded on plasmids [207]. Thioredoxins are known to act as antioxidants under oxidative stress conditions [135], but the significance of their role in fermented vegetable environments is not clear. Hyperosmotic stress during the production of such fermented foods may induce oxidative stress as reported in animal cells [208]. Group C is mainly composed of strains isolated from wine or grape environments in Spain, Italy, and Japan. Interestingly, genes involved in malolactic fermentation (MLF) are shared among many of them with high nucleotide identity. MLF is a decarboxylation process converting dicarboxylic malic acid into monocarboxylic lactic acid, which is exploited in wine production since it softens the acidity of wine [209]. It also confers benefits to organisms living in a harsh condition of wine because it can ameliorate acidic stress by reducing malic acid that is abundant in wine. In addition, some strains in this group share integrative and conjugative elements (ICEs) encoding genes for heavy metal resistance, phage resistance as well as oxidative resistance. These ICEs were also described in Chapter 2, in which *L. hokkaidonensis* and *L. vini* shared a large portion of the ICE components (Figure 2.4). By taking advantage of abounding genomic data, I newly found an ICE harbored by *L. uvarum* that exhibited high nucleotide identity with both *L. hokkaidonensis* and *L. vini* (Figure 4.11). As *L. hokkaidonensis* occupies a distinct niche from wine isolates, it is still difficult to presume the direct transfer between them. In constrast, *L. uvarum* and *L. vini* were both isolated from grape must in Spain, making the transfer of ICE between the two strains quite conceivable. Also, another strain of *L. vini* isolated in Brazil (strain JP7.8.9, GCA_000255515.2) does not harbor ICEs, suggesting the recent acquisition of these ICEs.

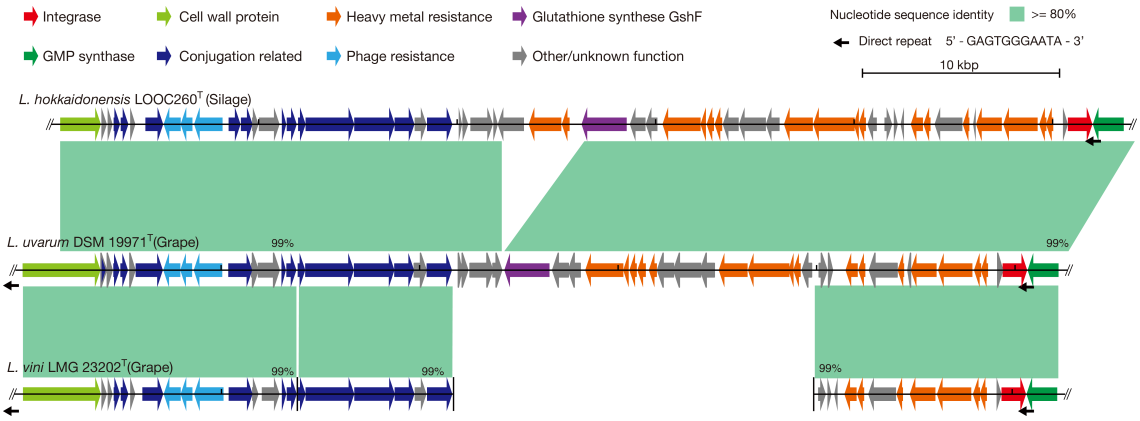


Figure 4.11. Integrated and conjugative elements from *L. hokkaidonensis* LOOC260^T, *L. uvarum* DSM 19971^T, and *L. vini* LMG 23202^T. Green and red correspond to the nucleotide identity based on BLASTN alignments and the numbers indicate the identity. Small black arrows represent direct repeat sequences flanking the element.

Horizontal acquisition of stress resistance genes is postulated in many studies [210]. This analysis exemplified the hypothesis by exploiting abundance of genomic data including ones reported from different research groups, and once again highlighted the role of mobile genetic elements as DNA vehicles especially in the stressful conditions. Of note, the strategy I took here did not consider the transfer between closely related species or small-scale transfer involving only a few genes. This might be improved by combining the *de novo* prediction method based on genomic signatures like GC content or codon frequency. In addition, phylogenetic analysis will further support the validity of gene transfer detected by this strategy.

4.4. Conclusions

I assessed 1,421 LAB genomes from 191 species, and archived them as a curated genome repository referred to as DAGA. Correct taxonomic names were assigned for 155 mislabeled or unidentified genomes and 38 genomes were marked as ‘poor quality’. Even genomes for type strains contain disqualified data due to possible misidentification or contamination. DAGA will improve the accessibility and reusability of LAB genome resources. The annotation and submission pipeline DFAST developed in this study will help researchers to deal with large amounts of emerging sequence data, thereby accelerating studies to further understanding of LAB on the basis of the genomic data.

By exploiting the data deposited in DAGA, I found previously unreported intraspecific genetic variation within three species, *Lactobacillus gasseri*, *Lactobacillus jensenii*, and *Lactobacillus vaginalis*, which might correspond to the subspecific level differentiation and deserve further characterization for their taxonomic validity. In addition, gene transfer analysis revealed the niche-specific dissemination of stress resistance genes among LAB genomes.

Through the assessment of the genomic data, the effectiveness of ANI in species classification and identification was demonstrated. Not limited to LAB, the use of ANI is widely spreading as evidence to describe new species. NCBI has started to use ANI to correct mislabeled entries in the GenBank database and has a plan to incorporate taxonomic validation in the early stage of the submission pipeline. It is also proposed that genomes of type strains should be sequenced when describing new bacterial species. This study took the initiative in establishing such a new era of microbial classification system.

Chapter 5

5. Conclusions and Perspective

In this study, I conducted three researches aiming to reveal the diverse characteristics and the evolutionary background of lactic acid bacteria (LAB) from a genomic perspective. The first one is the genome analysis of psychrotolerant LAB, *Lactobacillus hokkaidonensis*. The next one is the comparative analysis of fructophilic LAB, *Fructobacillus*, that cannot ferment glucose but fructose under anaerobic conditions. The last one describes the development of the genome annotation pipeline and its archive, DDBJ First Annotation and Submission Tool (DFAST) and DFAST Archive of Genome Annotation (DAGA), as well as demonstrative analyses utilizing hundreds of genomes deposited in DAGA.

In chapter 2, the genome sequence and analysis of *L. hokkaidonensis* LOOC260^{T} was presented. The complete genome derived by taking advantage of the third-generation sequencing platform realized the genome-wide understanding of mobile genetic elements, such as prophages, an integrative and conjugative element (ICE), and a conjugative plasmid. In particular, the ICEs found here encode stress resistance genes and are shared among LAB strains isolated from plants, suggesting their significance in adaptation to plant-associated environmental niches. Later, it was revealed by the analysis conducted in chapter 4 that the ICEs were shared among more LAB strains. Transporters for compatible solutes and the glutathione biosynthesis protein were identified as unique characteristics of this species that may contribute to the psychrotolerance mechanism. To confirm the findings of this study, temperature dependence of the gene expression profile is now under analysis using RNA-seq, and the preliminary analysis shows that several of these genes are induced at low temperature.

In chapter 3, the comparative genomics of the genera *Fructobacillus* and *Leuconostoc* was described. *Fructobacillus* spp. have smaller numbers of CDS in smaller genomes than *Leuconostoc* spp., which is due to specific gene loss of carbohydrate metabolic system. By this analysis, the general trend of reductive evolution in the fructose-rich environments was clearly revealed. The fructophilic property of *Fructobacillus* is attributable to cellular redox imbalance, which is also analogous to the preference for pentoses over glucose of *L. hokkaidonensis*. The characteristics and mechanisms to utilize sugars other than glucose may extend the potential for biotechnological application of LAB. This study gave insights into the linkage between physiological and biochemical characteristics of LAB and environmental factors in their habitats.

In chapter 4, I developed DAGA, in which I assessed LAB genomes obtained from both public sequence databases of DDBJ/ENA/GenBank and Sequence Read Archive (SRA). As a result, curated names inferred from average nucleotide identity (ANI) were assigned to 155 mislabeled or unidentified genomes, and 38 genomes were disqualified as 'poor quality genomes'. Through the development of DAGA, the effectiveness of ANI in bacterial classification and identification was demonstrated. This study took the initiative in establishing the new era of microbial classification system based on the whole genome information. Currently, DAGA stores 1,421 genomes covering 191 species of the family *Lactobacillaceae*. The reliable genomic information provided by DAGA will improve both accessibility and reusability of public sequence data for LAB. In addition, by leveraging the large dataset of DAGA, I revealed the previously unreported intraspecific diversity in *L. gasseri*, *L. jensenii*, and *L. vaginalis* and niche-specific dissemination of genes related to stress resistance. I also constructed curated reference database for LAB and developed DFAST as a web-based genome analysis platform. I have a plan to extend the scope of DAGA and DFAST to other groups of LAB as well as organisms other than LAB.

This study shed new light on the diversity and evolutionary background of LAB. LAB is a diverse and heterogeneous group that includes many organisms showing variety of characteristics such as a fructophilic species *L. kunkeei*, a coccoid lactobacillus like *L. dextrinicus*, motile LAB harboring flagellum systems, and strains have a potential to produce useful secondary metabolites. The abundance of genomic data will enable the unprecedented level of finer-grained analysis for such atypical strains, not limited to reference strains or industrially exploited strains, and open up a new horizon for understanding this organism. In the near future, comprehensive view of individual genomes for any kind of organisms would become available. The findings obtained and methods developed in this study will be applicable to expand our genomic view on many of microorganisms yet to be elucidated.

References

1. Winkler, H. *Verbreitung und ursache der parthenogenesis im pflanzen-und tierreiche*. (1920).
2. Suzuki, D. T. & Griffiths, A. J. F. DNA: The genetic material. in *An introduction to genetic analysis*. (eds. Miller, J. H. & Lewontin, R. C.) (W.H. Freeman and Company, 2000).
3. Saitou, N. *Introduction to Evolutionary Genomics*. 7–13 (Springer Science & Business Media, 2014).
4. Watson, J. D. & Crick, F. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* **171,** 737–738 (1953).
5. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A.* **74,** 560–564 (1977).
6. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94,** 441–448 (1975).
7. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **74,** 5463–5467 (1977).
8. Hutchison, C. A. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res.* **35,** 6227–6237 (2007).
9. Lander, E. S. *et al*. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
10. Venter, J. C. *et al*. The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
11. Karger, B. L. & Guttman, A. DNA Sequencing by Capillary Electrophoresis. *Electrophoresis* **30 Suppl 1,** S196–202 (2009).
12. Fleischmann, R. D. *et al*. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269,** 496–512 (1995).
13. Blattner, F. R. *et al*. The complete genome sequence of *Escherichia coli* K-12. *Science* **277,** 1453–1462 (1997).
14. Kunst, F. *et al*. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390,** 249–256 (1997).
15. Goffeau, A. *et al*. Life with 6000 genes. *Science* **274,** 546–563–7 (1996).
16. Land, M. *et al*. Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics* **15,** 141–161 (2015).

17. Metzker, M. L. Sequencing technologies — the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2009).

18. Vincent, A. T., Derome, N., Boyle, B., Culley, A. I. & Charette, S. J. Next-generation sequencing (NGS) in the microbiological world: How to make the most of your money. *J. Microbiol. Methods* S0167–7012(16)30031–8 (2016).

19. Reddy, T. B. K. *et al*. The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Res.* **43,** D1099–D1106 (2015).

20. Kyrpides, N. C. *et al*. Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains. *PLoS Biol.* **12,** e1001920 (2014).

21. Eid, J. *et al*. Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–138 (2009).

22. Stoddart, D., Heron, A. J., Mikhailova, E., Maglia, G. & Bayley, H. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 7702–7707 (2009).

23. Bertelli, C. & Greub, G. Rapid bacterial genome sequencing: methods and applications in clinical microbiology. *Clin. Microbiol. Infect.* **19,** 803–813 (2013).

24. Tettelin, H. *et al*. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U.S.A.* **102,** 13950–13955 (2005).

25. Rouli, L., Merhej, V., Fournier, P.-E. & Raoult, D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect.* **7,** 72–85 (2015).

26. Kaas, R. S., Friis, C., Ussery, D. W. & Aarestrup, F. M. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. *BMC Genomics* **13,** 577 (2012).

27. Janda, J. M. & Abbott, S. L. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J. Clin. Microbiol.* **45,** 2761–2764 (2007).

28. Delsuc, F., Brinkmann, H. & Philippe, H. Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* **6,** 361–375 (2005).

29. Wu, M. & Eisen, J. A. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.* **9,** R151 (2008).

30. Konstantinidis, K. T. & Tiedje, J. M. Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 2567–2572 (2005).

31. Goris, J. *et al*. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol*. **57,** 81–91 (2007).

32. Richter, M. & Rosselló-Mora, R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. U.S.A*. **106,** 19126–19131 (2009).

33. Auch, A. F., Jan, von, M., Klenk, H.-P. & Göker, M. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand. Genomic Sci*. **2,** 117–134 (2010).

34. Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P. & Göker, M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* **14,** 60 (2013).

35. Wayne, L. G. & Brenner, D. J. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int. J. Syst. Evol. Microbiol*. **37,** 463–464 (1987).

36. Olofsson, T. C., Alsterfjord, M., Nilson, B., Butler, E. & Vásquez, A. *Lactobacillus apinorum* sp. nov., *Lactobacillus mellifer* sp. nov., *Lactobacillus mellis* sp. nov., *Lactobacillus melliventris* sp. nov., *Lactobacillus kimbladii* sp. nov., *Lactobacillus helsingborgensis* sp. nov. and *Lactobacillus kullabergensis* sp. nov., isolated from the honey stomach of the honeybee *Apis mellifera*. *Int. J. Syst. Evol. Microbiol*. **64,** 3109–3119 (2014).

37. Puertas, A. I. *et al*. *Lactobacillus sicerae* sp. nov., a lactic acid bacterium isolated from Spanish natural cider. *Int. J. Syst. Evol. Microbiol*. **64,** 2949–2955 (2014).

38. Mao, Y., Chen, M. & Horvath, P. *Lactobacillus herbarum* sp. nov., a species related to *Lactobacillus plantarum*. *Int. J. Syst. Evol. Microbiol*. **65,** 4682–4688 (2015).

39. Lee, I., Kim, Y. O., Park, S.-C. & Chun, J. OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *Int. J. Syst. Evol. Microbiol*. **66,** 1100–1103 (2016).

40. Varghese, N. J. *et al*. Microbial species delineation using whole genome sequences. *Nucleic Acids Res*. **43,** 6761–6771 (2015).

41. Chan, J. Z.-M., Halachev, M. R., Loman, N. J., Constantinidou, C. & Pallen, M. J. Defining bacterial species in the genomic era: insights from the genus *Acinetobacter*. *BMC Microbiol*. **12,** 302 (2012).

42. Breitbart, M. *et al*. Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U.S.A*. **99,** 14250–14255 (2002).

43. Venter, J. C. *et al*. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304,** 66–74 (2004).

44. Tyson, G. W. *et al*. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428,** 37–43 (2004).

45. Sharpton, T. J. An introduction to the analysis of shotgun metagenomic data. *Front. Plant Sci.* **5,** 209 (2014).

46. Wang, W.-L. *et al*. Application of metagenomics in the human gut microbiome. *World J. Gastroenterol.* **21,** 803–814 (2015).

47. Brown, C. T. *et al*. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523,** 208–211 (2015).

48. Hug, L. A. *et al*. A new view of the tree of life. *Nat. Microbiol.* **48,** 1–6 (2016).

49. Apweiler, R., Bairoch, A. & Wu, C. H. Protein sequence databases. *Curr. Opin. Chem. Biol.* **8,** 76–80 (2004).

50. Strasser, B. J. The experimenter's museum: GenBank, natural history, and the moral economies of biomedicine. *Isis* **102,** 60–96 (2011).

51. Strasser, B. J. Genetics. GenBank--Natural history in the 21st Century? *Science* **322,** 537–538 (2008).

52. Cochrane, G., Karsch-Mizrachi, I., Takagi, T., International Nucleotide Sequence Database Collaboration. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.* **44,** D48–50 (2016).

53. Kodama, Y., Shumway, M., Leinonen, R., International Nucleotide Sequence Database Collaboration. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.* **40,** D54–6 (2012).

54. Barrett, T. *et al*. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.* **40,** D57–63 (2012).

55. Federhen, S. *et al*. Meeting report: GenBank microbial genomic taxonomy workshop(12–13 May, 2015). *Stand. Genomic Sci.* **11,** 15 (2016).

56. O'Leary, N. A. *et al*. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44,** D733–45 (2016).

57. Tatusova, T. *et al*. Update on RefSeq microbial genomes resources. *Nucleic Acids Res.* **43,** D599–605 (2015).

58. Kitts, P. A. *et al*. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.* **44,** D73–80 (2016).

59. Land, M. L. *et al*. Quality scores for 32,000 genomes. *Stand. Genomic Sci.* **9,** 20 (2014).

60. Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K. & Tolstoy, I. RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res*. **42,** D553–9 (2014).

61. Figueras, M.-J., Beaz-Hidalgo, R., Hossain, M. J. & Liles, M. R. Taxonomic affiliation of new genomes should be verified using average nucleotide identity and multilocus phylogenetic analysis. *Genome Announc*. **2,** e00927–14 (2014).

62. Beaz-Hidalgo, R., Hossain, M. J., Liles, M. R. & Figueras, M.-J. Strategies to avoid wrongly labelled genomes using as example the detected wrong taxonomic affiliation for aeromonas genomes in the GenBank database. *PLoS ONE* **10,** e0115813 (2015).

63. Salvetti, E., Torriani, S. & Felis, G. E. The genus *Lactobacillus*: a taxonomic update. *Probiotics Antimicrob. Proteins* **4,** 217–226 (2012).

64. Makarova, K. *et al*. Comparative genomics of the lactic acid bacteria. *Proc. Natl. Acad. Sci. U.S.A.* **103,** 15611–15616 (2006).

65. Mazzoli, R., Bosco, F., Mizrahi, I., Bayer, E. A. & Pessione, E. Towards lactic acid bacteria-based biorefineries. *Biotechnol. Adv*. **32,** 1216–1236 (2014).

66. Quinto, E. J. *et al*. Probiotic lactic acid bacteria: a review. *FNS* **5,** 1765–1775 (2014).

67. Cano-Garrido, O., Seras-Franzoso, J. & Garcia-Fruitós, E. Lactic acid bacteria: reviewing the potential of a promising delivery live vector for biomedical purposes. *Microb. Cell Fact*. **14,** 137 (2015).

68. de Vos, W. M. Systems solutions by lactic acid bacteria: from paradigms to practice. *Microb. Cell Fact*. **10 Suppl 1,** S2 (2011).

69. Swain, M. R., Anandharaj, M., Ray, R. C. & Parveen Rani, R. Fermented fruits and vegetables of Asia: a potential source of probiotics. *Biotechnol. Res. Int*. **2014,** 250424 (2014).

70. Zhang, Z.-G., Ye, Z.-Q., Yu, L. & Shi, P. Phylogenomic reconstruction of lactic acid bacteria: an update. *BMC Evol. Biol*. **11,** 1 (2011).

71. Vandamme, P., De Bruyne, K. & Pot, B. Phylogenetics and systematics. in *Lactic acid bacteria: biodiversity and taxonomy* (eds. Holzapfel, W. H. & Wood, B. J. B.) 31–44 (John Wiley & Sons, Ltd., 2014).

72. Mattarelli, P. & Biavati, B. The genera *Bifidobacterium*, *Parascardovia* and *Scardovia*. in *Lactic Acid Bacteria: Biodiversity and Taxonomy* (eds. Holzapfel, W. H. & Wood, B. J. B.) 509–541 (John Wiley & Sons, Ltd., 2014).

73. Hammes, W. P. & Hertel, C. *Bergey's Manual of Systematic Bacteriology: Volume 3: The Firmicutes*. 465–511 (Springer Science & Business Media, 2009).

74. Franz, C. M. A. P. *et al*. The genus *Pediococcus*. in *Lactic Acid Bacteria: Biodiversity and Taxonomy* (eds. Holzapfel, W. H. & Wood, B. J. B.) 359–376 (John Wiley & Sons, 2014).

75. Zheng, J., Ruan, L., Sun, M. & Gänzle, M. A genomic view of lactobacilli and pediococci demonstrates that phylogeny matches ecology and physiology. *Appl. Environ. Microbiol.* **81,** 7233–7243 (2015).

76. Haakensen, M., Dobson, C. M., Hill, J. E. & Ziola, B. Reclassification of *Pediococcus dextrinicus* (Coster and White 1964) back 1978 (Approved Lists 1980) as *Lactobacillus dextrinicus* comb. nov., and emended description of the genus *Lactobacillus*. *Int. J. Syst. Evol. Microbiol*. **59,** 615–621 (2009).

77. Collins, M. D., Williams, A. M. & Wallbanks, S. The phylogeny of *Aerococcus* and *Pediococcus* as determined by 16S rRNA sequence analysis: description of *Tetragenococcus* gen. nov. *FEMS Microbiol. Lett*. **58,** 255–262 (1990).

78. Dicks, L. M., Dellaglio, F. & Collins, M. D. Proposal to reclassify *Leuconostoc oenos* as *Oenococcus oeni* [corrig.] gen. nov., comb. nov.. *Int. J. Syst. Bacteriol*. **45,** 395–397 (1995).

79. Endo, A. & Okada, S. Reclassification of the genus *Leuconostoc* and proposals of *Fructobacillus fructosus* gen. nov., comb. nov., *Fructobacillus durionis* comb. nov., *Fructobacillus ficulneus* comb. nov. and *Fructobacillus pseudoficulneus* comb. nov. *Int. J. Syst. Evol. Microbiol*. **58,** 2195–2205 (2008).

80. Pot, B. & Tsakalidou, E. Taxonomy and Metabolism of *Lactobacillus*. in *Lactobacillus molecular biology: from genomics to probiotics* (eds. Ljungh, Å. & Wadström, T.) 3–59 (Caister Academic Press, 2009).

81. Smokvina, T. *et al*. *Lactobacillus paracasei* comparative genomics: towards species pan-genome definition and exploitation of diversity. *PLoS ONE* **8,** e68731 (2013).

82. Bolotin, A. *et al*. The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. lactis IL1403. *Genome Res*. **11,** 731–753 (2001).

83. Kleerebezem, M. *et al*. Complete genome sequence of *Lactobacillus plantarum* WCFS1. *Proc. Natl. Acad. Sci. U.S.A*. **100,** 1990–1995 (2003).

84. Douillard, F. P. & de Vos, W. M. Functional genomics of lactic acid bacteria: from food to health. *Microb. Cell Fact*. **13 Suppl 1,** S8 (2014).

85. Canchaya, C., Claesson, M. J., Fitzgerald, G. F., van Sinderen, D. & O'Toole, P. W. Diversity of the genus *Lactobacillus* revealed by comparative genomics of five species. *Microbiology* **152,** 3185–3196 (2006).

86. Kant, R., Blom, J., Palva, A., Siezen, R. J. & de Vos, W. M. Comparative genomics of *Lactobacillus*. *Microb. Biotechnol.* **4,** 323–332 (2010).

87. Douillard, F. P. *et al*. Comparative genomic and functional analysis of 100 *Lactobacillus rhamnosus* strains and their comparison with strain GG. *PLoS Genet.* **9,** e1003683 (2013).

88. Sternes, P. R. & Borneman, A. R. Consensus pan-genome assembly of the specialised wine bacterium *Oenococcus oeni*. *BMC Genomics* **17,** 308 (2016).

89. Sun, Z. *et al*. Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. *Nat. Commun.* **6,** 8322 (2015).

90. Tohno, M. *et al*. *Lactobacillus hokkaidonensis* sp. nov., isolated from subarctic timothy grass (*Phleum pratense* L.) silage. *Int. J. Syst. Evol. Microbiol.* **63,** 2526–2531 (2013).

91. Bull, M. J., Marchesi, J. R., Vandamme, P., Plummer, S. & Mahenthiralingam, E. Minimum taxonomic criteria for bacterial genome sequence depositions and announcements. *J. Microbiol. Methods* **89,** 18–21 (2012).

92. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25,** 1043–1055 (2015).

93. Pot, B. *et al*. The genus *Lactobacillus*. in *Lactic Acid Bacteria: Biodiversity and Taxonomy* (eds. Holzapfel, W. E. & Wood, B. J. B.) 249–353 (John Wiley & Sons, Ltd., 2014).

94. Okada, S., Suzuki, Y. & Kozaki, M. New heterofermentative *Lactobacillus* species with meso-diaminopimelic acid in peptidoglycan, *Lactobacillus vaccinostercus* Kozaki and Okada sp. nov. *J. Gen. Appl. Microbiol.* **25,** 215–221 (1979).

95. Kleynmans, U., Heinzl, H. & Hammes, W. P. *Lactobacillus suebicus* sp. nov., an obligately heterofermentative *Lactobacillus* species isolated from fruit mashes. *Syst. Appl. Microbiol.* **11,** 267–271 (1989).

96. Koort, J. *et al*. *Lactobacillus oligofermentans* sp. nov., associated with spoilage of modified-atmosphere-packaged poultry products. *Appl. Environ. Microbiol.* **71,** 4400–4406 (2005).

97. Gu, C. T., Li, C. Y., Yang, L. J. & Huo, G. C. *Lactobacillus mudanjiangensis* sp. nov., *Lactobacillus songhuajiangensis* sp. nov. and *Lactobacillus nenjiangensis* sp. nov., isolated from Chinese traditional pickle and sourdough. *Int. J. Syst. Evol. Microbiol.* **63,** 4698–4706 (2013).

98. Kajitani, R. *et al*. Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **24,** 1384–1395 (2014).

99. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

100. Lin, I. H., Chen, K. B., Lin, Y. H. & Chang, C. H. Automated Prediction Of Bacterial Replication Origin (APBRO). in *Proceedings of the 4th Asia Pacific Bioinformatics Conference*, Taipei, Taiwan. (2006).

101. Sugawara, H., Ohyama, A., Mori, H. & Kurokawa, K. Microbial genome annotation pipeline (MiGAP) for diverse users. in *Proceedings of the 20th International Conference on Genome Informatics*, Yokohama, Japan. S–001–1–2 (2009).

102. Noguchi, H., Taniguchi, T. & Itoh, T. MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res*. **15,** 387–396 (2008).

103. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*. **25,** 955–964 (1997).

104. Lagesen, K. *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res*. **35,** 3100–3108 (2007).

105. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res*. **35,** W182–5 (2007).

106. Varani, A. M., Siguier, P., Gourbeyre, E., Charneau, V. & Chandler, M. ISsaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biol*. **12,** R30 (2011).

107. Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. & Wishart, D. S. PHAST: a fast phage search tool. *Nucleic Acids Res*. **39,** W347–52 (2011).

108. Grissa, I., Vergnaud, G. & Pourcel, C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res*. **35,** W52–7 (2007).

109. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. **32,** 1792–1797 (2004).

110. Tamura, K. *et al*. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol*. **28,** 2731–2739 (2011).

111. Krzywinski, M. *et al*. Circos: an information aesthetic for comparative genomics. *Genome Res*. **19,** 1639–1645 (2009).

112. Pritchard, L., White, J. A., Birch, P. R. J. & Toth, I. K. GenomeDiagram: a python package for the visualization of large-scale genomic data. *Bioinformatics* **22,** 616–617 (2006).

113. Chin, C.-S. *et al*. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10,** 563–569 (2013).

114. Molenaar, D. *et al*. Exploring *Lactobacillus plantarum* genome diversity by using microarrays. *J. Bacteriol*. **187,** 6119–6127 (2005).

115. Cai, H., Thompson, R., Budinich, M. F., Broadbent, J. R. & Steele, J. L. Genome sequence and comparative genome analysis of *Lactobacillus casei*: insights into their niche-associated evolution. *Genome Biol. Evol*. **1,** 239–257 (2009).

116. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res*. **34,** D32–6 (2006).

117. van Kranenburg, R. *et al*. Functional analysis of three plasmids from *Lactobacillus plantarum*. *Appl. Environ. Microbiol*. **71,** 1223–1230 (2005).

118. Fukao, M. *et al*. Genomic analysis by deep sequencing of the probiotic *Lactobacillus brevis* KB290 harboring nine plasmids reveals genomic stability. *PLoS ONE* **8,** e60521 (2013).

119. Tohno, M. *et al*. *Lactobacillus oryzae* sp. nov., isolated from fermented rice grain (*Oryza sativa* L. subsp. *japonica*). *Int. J. Syst. Evol. Microbiol*. **63,** 2957–2962 (2013).

120. Tanizawa, Y. *et al*. Draft Genome Sequence of *Lactobacillus oryzae* Strain SG293^{T}. *Genome Announc*. **2,** e00861–14 (2014).

121. Wozniak, R. A. F. & Waldor, M. K. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nat. Rev. Microbiol*. **8,** 552–563 (2010).

122. Burrus, V. & Waldor, M. K. Shaping bacterial genomes with integrative and conjugative elements. *Res. Microbiol*. **155,** 376–386 (2004).

123. Devirgiliis, C., Coppola, D., Barile, S., Colonna, B. & Perozzi, G. Characterization of the Tn*916* conjugative transposon in a food-borne strain of *Lactobacillus paracasei*. *Appl. Environ. Microbiol*. **75,** 3866–3871 (2009).

124. Raftis, E. J., Forde, B. M., Claesson, M. J. & O'Toole, P. W. Unusual genome complexity in *Lactobacillus salivarius* JCM1046. *BMC Genomics* **15,** 771 (2014).

125. Bi, D. *et al*. ICEberg: a web-based resource for integrative and conjugative elements found in Bacteria. *Nucleic Acids Res*. **40,** D621–6 (2012).

126. Roberts, A. P. *et al*. Revised nomenclature for transposable genetic elements. *Plasmid* **60,** 167–173 (2008).

127. Song, L., Pan, Y., Chen, S. & Zhang, X. Structural characteristics of genomic islands associated with GMP synthases as integration hotspot among sequenced microbial genomes. *Comput. Biol. Chem*. **36,** 62–70 (2012).

128. van de Guchte, M. *et al*. Stress responses in lactic acid bacteria. *Antonie van Leeuwenhoek* **82,** 187–216 (2002).

129. Chattopadhyay, M. K. Mechanism of bacterial adaptation to low temperature. *J. Biosci*. **31,** 157–165 (2006).

130. Barria, C., Malecki, M. & Arraiano, C. M. Bacterial adaptation to cold. *Microbiology* **159,** 2437–2443 (2013).

131. Hoffmann, T. & Bremer, E. Protection of *Bacillus subtilis* against cold stress via compatible-solute acquisition. *J. Bacteriol*. **193,** 1552–1562 (2011).

132. Angelidis, A. S. & Smith, G. M. Role of the glycine betaine and carnitine transporters in adaptation of *Listeria monocytogenes* to chill stress in defined medium. *Appl. Environ. Microbiol*. **69,** 7492–7498 (2003).

133. Annamalai, T. & Venkitanarayanan, K. Role of *proP* and *proU* in betaine uptake by *Yersinia enterocolitica* under cold and osmotic stress conditions. *Appl. Environ. Microbiol*. **75,** 1471–1477 (2009).

134. Chaillou, S. *et al*. The complete genome sequence of the meat-borne lactic acid bacterium *Lactobacillus sakei* 23K. *Nat. Biotechnol*. **23,** 1527–1533 (2005).

135. Pophaly, S. D., Singh, R., Pophaly, S. D., Kaushik, J. K. & Tomar, S. K. Current status and emerging role of glutathione in food grade lactic acid bacteria. *Microb. Cell Fact*. **11,** 114 (2012).

136. Zhang, J. *et al*. Glutathione protects *Lactobacillus sanfranciscensis* against freeze-thawing, freeze-drying, and cold treatment. *Appl. Environ. Microbiol*. **76,** 2989–2996 (2010).

137. Zhang, J., Li, Y., Chen, W., Du, G.-C. & Chen, J. Glutathione improves the cold resistance of *Lactobacillus sanfranciscensis* by physiological regulation. *Food Microbiol*. **31,** 285–292 (2012).

138. Heinl, S. *et al*. Insights into the completely annotated genome of *Lactobacillus buchneri* CD034, a strain isolated from stable grass silage. *J. Biotechnol*. **161,** 153–166 (2012).

139. Hayashi, T., Okada, S. & Kozaki, M. Effects of some potential electron acceptors on glucose as a sole energy source for the growth of *Lactobacillus vaccinostercus*. *J. Gen. Appl. Microbiol*. **28,** 87–94 (1982).

140. Warriner, K. & Morris, J. G. The effects of aeration on the bioreductive abilities of some heterofermentative lactic acid bacteria. *Lett. Appl. Microbiol*. **20,** 323–327 (1995).

141. Endo, A., Tanaka, N., Oikawa, Y., Okada, S. & Dicks, L. Fructophilic characteristics of *Fructobacillus* spp. may be due to the absence of an alcohol/acetaldehyde dehydrogenase gene (*adhE*). *Curr. Microbiol.* **68,** 531–535 (2014).

142. Weinberg, Z. G. & Muck, R. E. New trends and opportunities in the development and use of inoculants for silage. *FEMS Microbiol. Rev.* **19,** 53–68 (1996).

143. Danner, H., Holzer, M., Mayrhuber, E. & Braun, R. Acetic acid increases stability of silage under aerobic conditions. *Appl. Environ. Microbiol.* **69,** 562–567 (2003).

144. Lee, W.-H., Kim, M.-D., Jin, Y.-S. & Seo, J.-H. Engineering of NADPH regenerators in *Escherichia coli* for enhanced biotransformation. *Appl. Microbiol. Biotechnol.* **97,** 2761–2772 (2013).

145. Takeno, S., Murata, R., Kobayashi, R., Mitsuhashi, S. & Ikeda, M. Engineering of *Corynebacterium glutamicum* with an NADPH-generating glycolytic pathway for L-Lysine production. *Appl. Environ. Microbiol.* **76,** 7154–7160 (2010).

146. Boyd, D. A., Cvitkovitch, D. G. & Hamilton, I. R. Sequence, expression, and function of the gene for the nonphosphorylating, NADP-dependent glyceraldehyde-3-phosphate dehydrogenase of *Streptococcus mutans*. *J. Bacteriol.* **177,** 2622–2627 (1995).

147. Slattery, L., O'Callaghan, J., Fitzgerald, G. F., Beresford, T. & Ross, R. P. Invited review: *Lactobacillus helveticus*—A thermophilic dairy starter related to gut bacteria. *J. Dairy Sci.* **93,** 4435–4454 (2010).

148. Nomura, M., Kobayashi, M., Narita, T., Kimoto-Nira, H. & Okamoto, T. Phenotypic and molecular characterization of *Lactococcus lactis* from milk and plants. *J. Appl. Microbiol.* **101,** 396–405 (2006).

149. Hammes, W. P. & Hertel, C. The Genera *Lactobacillus* and *Carnobacterium*. in *The Prokaryotes: Vol. 4: Bacteria: Firmicutes, Cyanobacteria* (eds. Falkow, S., Rosenberg, E., Schleifer, K.-H. & Stackebrandt, E.) 320–403 (Springer Science & Business Media, 2006).

150. Endo, A., Futagawa-Endo, Y. & Dicks, L. M. T. Isolation and characterization of fructophilic lactic acid bacteria from fructose-rich niches. *Syst. Appl. Microbiol.* **32,** 593–600 (2009).

151. Endo, A. & Salminen, S. Honeybees and beehives are rich sources for fructophilic lactic acid bacteria. *Syst. Appl. Microbiol.* **36,** 444–448 (2013).

152. Endo, A. & Dicks, L. M. T. The genus *Fructobacillus*. in *Lactic Acid Bacteria: Biodiversity and Taxonomy* (eds. Holzapfel, W. H. & Wood, B. J. B.) 381–390 (John Wiley & Sons, Ltd., 2014).

153. Endo, A. *et al*. *Fructobacillus tropaeoli* sp. nov., a fructophilic lactic acid bacterium isolated from a flower. *Int*. *J*. *Syst*. *Evol*. *Microbiol*. **61,** 898–902 (2011).

154. Mendes-Soares, H., Suzuki, H., Hickey, R. J. & Forney, L. J. Comparative functional genomics of *Lactobacillus* spp. reveals possible mechanisms for specialization of vaginal lactobacilli to their environment. *J*. *Bacteriol*. **196,** 1458–1470 (2014).

155. van de Guchte, M. *et al*. The complete genome sequence of *Lactobacillus bulgaricus* reveals extensive and ongoing reductive evolution. *Proc*. *Natl*. *Acad*. *Sci*. *U.S.A*. **103,** 9274–9279 (2006).

156. Hols, P. *et al*. New insights in the molecular biology and physiology of *Streptococcus thermophilus* revealed by comparative genomics. *FEMS Microbiol*. *Rev*. **29,** 435–463 (2005).

157. Endo, A. & Okada, S. Monitoring the lactic acid bacterial diversity during shochu fermentation by PCR-denaturing gradient gel electrophoresis. *J*. *Biosci*. *Bioeng*. **99,** 216–221 (2005).

158. Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res*. **18,** 821–829 (2008).

159. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*. **28,** 33–36 (2000).

160. Kyrpides, N. C. *et al*. Genomic Encyclopedia of Type Strains, Phase I: The one thousand microbial genomes (KMG-I) project. *Stand*. *Genomic Sci*. **9,** 1278–1284 (2014).

161. Contreras-Moreira, B. & Vinuesa, P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl*. *Environ*. *Microbiol*. **79,** 7696–7701 (2013).

162. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst*. *Biol*. **56,** 564–577 (2007).

163. Kück, P. & Longo, G. C. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Front*. *Zool*. **11,** 81 (2014).

164. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).

165. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27,** 1164–1165 (2011).

166. Endo, A. & Okada, S. *Lactobacillus satsumensis* sp. nov., isolated from mashes of shochu, a traditional Japanese distilled spirit made from fermented rice and other starchy materials. *Int*. *J*. *Syst*. *Evol*. *Microbiol*. **55,** 83–85 (2005).

167. Lukjancenko, O., Ussery, D. W. & Wassenaar, T. M. Comparative genomics of *Bifidobacterium*, *Lactobacillus* and related probiotic genera. *Microb. Ecol.* **63,** 651–673 (2011).

168. Rahkila, R., De Bruyne, K., Johansson, P., Vandamme, P. & Björkroth, J. Reclassification of *Leuconostoc gasicomitatum* as *Leuconostoc gelidum* subsp. *gasicomitatum* comb. nov., description of *Leuconostoc gelidum* subsp. *aenigmaticum* subsp. nov., designation of *Leuconostoc gelidum* subsp. *gelidum* subsp. nov. and emended description of *Leuconostoc gelidum*. *Int. J. Syst. Evol. Microbiol.* **64,** 1290–1295 (2014).

169. Björkroth, K. J. *et al*. Characterization of *Leuconostoc gasicomitatum* sp. nov., associated with spoiled raw tomato-marinated broiler meat strips packaged under modified-atmosphere conditions. *Appl. Environ. Microbiol.* **66,** 3764–3772 (2000).

170. Jääskeläinen, E. *et al*. Significance of heme-based respiration in meat spoilage caused by *Leuconostoc gasicomitatum*. *Appl. Environ. Microbiol.* **79,** 1078–1085 (2013).

171. Björkroth, J. & Holzapfel, W. Genera *Leuconostoc*, *Oenococcus* and *Weissella*. in *The Prokaryotes* (eds. Dworkin, M., Falkow, S., Rosenberg, E., Schleifer, K.-H. & Stackebrandt, E.) 267–319 (Springer US, 2006).

172. Arevalo-Villena, M., Bartowsky, E. J., Capone, D. & Sefton, M. A. Production of indole by wine-associated microorganisms under oenological conditions. *Food Microbiol.* **27,** 685–690 (2010).

173. Tamarit, D. *et al*. Functionally structured genomes in *Lactobacillus kunkeei* colonizing the honey crop and food products of honeybees and stingless bees. *Genome Biol. Evol.* **7,** 1455–1473 (2015).

174. Teixeira, J. S., McNeill, V. & Gänzle, M. G. Levansucrase and sucrose phoshorylase contribute to raffinose, stachyose, and verbascose metabolism by lactobacilli. *Food Microbiol.* **31,** 278–284 (2012).

175. Tieking, M., Ehrmann, M. A., Vogel, R. F. & Gänzle, M. G. Molecular and functional characterization of a levansucrase from the sourdough isolate *Lactobacillus sanfranciscensis* TMW 1.392. *Appl. Microbiol. Biotechnol.* **66,** 655–663 (2005).

176. Velázquez-Hernández, M. L. *et al*. *Gluconacetobacter diazotrophicus* levansucrase is involved in tolerance to NaCl, sucrose and desiccation, and in biofilm formation. *Arch. Microbiol.* **193,** 137–149 (2011).

177. Nieminen, T. T., Säde, E., Endo, A., Johansson, P. & Björkroth, J. The family *Leuconostocaceae*. in *The Prokaryotes* (eds. Rosenberg, E., DeLong, E. F., Lory, S., Stackebrandt, E. & Thompson, F. L.) 215–240 (Springer Berlin Heidelberg, 2014).

178. Nagy, A. *et al*. Identification and correction of abnormal, incomplete and mispredicted proteins in public databases. *BMC Bioinformatics* **9,** 353 (2008).

179. Nilsson, R. H. *et al*. Taxonomic reliability of DNA Sequences in public sequence databases: a fungal perspective. *PLoS ONE* **1,** e59 (2006).

180. Nakazato, T., Ohta, T. & Bono, H. Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive. *PLoS ONE* **8,** e77910 (2013).

181. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29,** 1072–1075 (2013).

182. Kim, M., Oh, H.-S., Park, S.-C. & Chun, J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int. J. Syst. Evol. Microbiol.* **64,** 346–351 (2014).

183. Kim, O.-S. *et al*. Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. *Int. J. Syst. Evol. Microbiol.* **62,** 716–721 (2012).

184. Mattarelli, P. *et al*. Recommended minimal standards for description of new taxa of the genera *Bifidobacterium*, *Lactobacillus* and related genera. *Int. J. Syst. Evol. Microbiol.* **64,** 1434–1451 (2014).

185. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068–2069 (2014).

186. Sugawara, H., Miyazaki, S., Gojobori, T. & Tateno, Y. DNA Data Bank of Japan dealing with large-scale data submission. *Nucleic Acids Res.* **27,** 25–28 (1999).

187. Uchiyama, I. MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. *Nucleic Acids Res.* **35,** D343–6 (2007).

188. Marchler-Bauer, A. *et al*. CDD: NCBI's conserved domain database. *Nucleic Acids Res.* **43,** D222–6 (2015).

189. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **216,** 403–410 (1990).

190. Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27,** 431–432 (2011).

191. Morita, H. *et al*. *Sharpea azabuensis* gen. nov., sp. nov., a Gram-positive, strictly anaerobic bacterium isolated from the faeces of thoroughbred horses. *Int. J. Syst. Evol. Microbiol.* **58,** 2682–2686 (2008).

192. Salvetti, E. *et al*. Reclassification of *Lactobacillus catenaformis* (Eggerth 1935) Moore and Holdeman 1970 and *Lactobacillus vitulinus* Sharpe et al. 1973 as *Eggerthia catenaformis* gen. nov., comb. nov. and *Kandleria vitulina* gen. nov., comb. nov., respectively. *Int. J. Syst. Evol. Microbiol*. **61,** 2520–2524 (2011).

193. Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K. & Tolstoy, I. RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res*. **43,** 3872 (2015).

194. Wattam, A. R. *et al*. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res*. **42,** D581–91 (2014).

195. Judicial Commission of the International Committee on Systematics of Bacteria. The type strain of *Lactobacillus casei* is ATCC 393, ATCC 334 cannot serve as the type because it represents a different taxon, the name *Lactobacillus paracasei* and its subspecies names are not rejected and the revival of the name '*Lactobacillus zeae*' contravenes Rules 51b (1) and (2) of the International Code of Nomenclature of Bacteria. Opinion 82. *Int. J. Syst. Evol. Microbiol*. **58,** 1764–1765 (2008).

196. Felis, G. E. & Pot, B. The family *Lactobacillaceae*. in (eds. Holzapfel, W. H. & Wood, B. J. B.) 245–247 (John Wiley & Sons, Ltd., 2014).

197. Wieme, A., Cleenwerck, I., Van Landschoot, A. & Vandamme, P. *Pediococcus lolii* DSM 19927^{T} and JCM 15055^{T} are strains of *Pediococcus acidilactici*. *Int. J. Syst. Evol. Microbiol*. **62,** 3105–3108 (2012).

198. Skerman, V., McGowan, V. & Sneath, P. Approved lists of bacterial names. *Int. J. Syst. Bacteriol*. **30,** 225–420 (1980).

199. Kitahara, M. Quality management of *Lactobacillus* strains in JCM. *Microbiol. Cul. Col*. **24,** 143–145 (2008).

200. Goto, N., Joyeux, A. & Lonvaud-Funel, A. Taxonomic Problem of the Type Strain of *Lactobacillus homohiochii*. *J. Brew. Soc. Jpn*. **89,** 643–646 (1994).

201. Yamamoto, H. S., Xu, Q. & Fichorova, R. N. Homeostatic properties of *Lactobacillus jensenii* engineered as a live vaginal anti-HIV microbicide. *BMC Microbiol*. **13,** 4 (2013).

202. Ravel, J. *et al*. Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci. U.S.A.* **108 Suppl 1,** 4680–4687 (2011).

203. Tanigawa, K. & Watanabe, K. Multilocus sequence typing reveals a novel subspeciation of *Lactobacillus delbrueckii*. *Microbiology* **157,** 727–738 (2011).

204. Adimpong, D. B. *et al*. *Lactobacillus delbrueckii* subsp. *jakobsenii* subsp. nov., isolated from dolo wort, an alcoholic fermented beverage in Burkina Faso. *Int. J. Syst. Evol. Microbiol*. **63,** 3720–3726 (2013).

205. Suzuki, K. 125th Anniversary Review: Microbiological Instability of Beer Caused by Spoilage Bacteria. *J. Inst. Brew*. **117,** 131–155 (2011).

206. Bergsveinson, J., Baecker, N., Pittet, V. & Ziola, B. Role of plasmids in *Lactobacillus brevis* BSO 464 hop tolerance and beer spoilage. *Appl. Environ. Microbiol*. **81,** 1234–1241 (2015).

207. Oh, H.-M. *et al*. Complete genome sequence analysis of *Leuconostoc kimchii* IMSNU 11154. *J. Bacteriol*. **192,** 3844–3845 (2010).

208. McCarthy, M. J., Baumber, J., Kass, P. H. & Meyers, S. A. Osmotic stress induces oxidative cell damage to rhesus macaque spermatozoa. *Biol. Reprod*. **82,** 644–651 (2010).

209. Liu, S.-Q. A review: malolactic fermentation in wine -- beyond deacidification. *J. Appl. Microbiol*. **92,** 589–601 (2002).

210. Shoeb, E. *et al*. Horizontal gene transfer of stress resistance genes through plasmid transport. *World J. Microb. Biot*. **28,** 1021–1025 (2012).

211. Guo F-B, Lin H, Huang J. A plot of G + C content against sequence length of 640 bacterial chromosomes shows the points are widely scattered in the upper triangular area. *Chromosome Res*. **17,** 359–364 (2009).

212. Haakensen, M. & Ziola, B. Identification of novel *horA*-harbouring bacteria capable of spoiling beer. *Can. J. Microbiol*. **54,** 321–325 (2008).

Acknowledgements

First of all, I am deeply indebted to my three supervisors, Prof. Toshihisa Takagi, Prof. Yasukazu Nakamura, and Prof. Masanori Arita for giving me the wonderful opportunity to carry out my research program here at the Graduate School of Frontier Sciences of the University of Tokyo. I am so grateful for their continuous guidance, meticulous suggestions, and perspective criticism, which helped me to expand my knowledge and scientific view. Without their precious support, it would not have been possible to conduct this research. My thanks also go to my dissertation committee members, Prof. Yutaka Suzuki, Assoc. Prof. Wataru Iwasaki, Prof. Masahira Hattori, and Prof. Ken Kurokawa reviewing my dissertation and giving unstinted suggestions for improvement.

I also appreciate my research collaborators, Dr. Masanori Tohno at the NARO Institute of Livestock and Grassland Science and Dr. Akihito Endo at Tokyo University of Agriculture, for introducing me to the study of lactic acid bacteria. Their extensive knowledge and experience in microbiology inspired me a lot, and the exciting discussions with them always provided me fruitful insights. Additionally, I would like to thank Dr. Eli Kaminuma and all the laboratory members for their sound advices, insightful suggestions, and also daily conversations that always made me feel relaxed.

Finally, I would like to express my deepest gratitude to my parents, my two grandmas, my wife Satomi, and my beloved daughter Maika for encouraging me and keeping me motivated throughout my doctoral program.