Full Paper Bioscience of Microbiota, Food and Health Vol. 35 (4), 173–184, 2016

DFAST and DAGA: web-based integrated genome annotation tools and resources

Yasuhiro TANIZAWA[1, 2], Takatomo FUJISAWA[2], Eli KAMINUMA[2], Yasukazu NAKAMURA[2] and Masanori ARITA[2, 3*]

[1]*Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561, Japan*

[2]*Center for Information Biology, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan*

[3]*RIKEN Center for Sustainable Resource Science, 1-7-22 Suehiro-cho, Tsurumi, Yokohama, Kanagawa 230-0045, Japan*

Received February 15, 2016; Accepted June 27, 2016; Published online in J-STAGE July 14, 2016

Quality assurance and correct taxonomic affiliation of data submitted to public sequence databases have been an everlasting problem. The DDBJ Fast Annotation and Submission Tool (DFAST) is a newly developed genome annotation pipeline with quality and taxonomy assessment tools. To enable annotation of ready-to-submit quality, we also constructed curated reference protein databases tailored for lactic acid bacteria. DFAST was developed so that all the procedures required for DDBJ submission could be done seamlessly online. The online workspace would be especially useful for users not familiar with bioinformatics skills. In addition, we have developed a genome repository, DFAST Archive of Genome Annotation (DAGA), which currently includes 1,421 genomes covering 179 species and 18 subspecies of two genera, *Lactobacillus* and *Pediococcus*, obtained from both DDBJ/ENA/GenBank and Sequence Read Archive (SRA). All the genomes deposited in DAGA were annotated consistently and assessed using DFAST. To assess the taxonomic position based on genomic sequence information, we used the average nucleotide identity (ANI), which showed high discriminative power to determine whether two given genomes belong to the same species. We corrected mislabeled or misidentified genomes in the public database and deposited the curated information in DAGA. The repository will improve the accessibility and reusability of genome resources for lactic acid bacteria. By exploiting the data deposited in DAGA, we found intraspecific subgroups in *Lactobacillus gasseri* and *Lactobacillus jensenii*, whose variation between subgroups is larger than the well-accepted ANI threshold of 95% to differentiate species. DFAST and DAGA are freely accessible at https://dfast.nig.ac.jp.

Key words: lactic acid bacteria, genome, annotation, database, *Lactobacillus*, *Pediococcus*

INTRODUCTION

Major scientific journals request that researchers deposit newly sequenced DNA in the International Nucleotide Sequence Database Collaboration (INSDC) [1]. DDBJ/ENA/GenBank are the core annotation databases, collecting publicly available DNA information with metadata. Recently, INSDC has also begun collecting raw sequences from the new-generation sequencing platforms for Sequence Read Archive (SRA) [2]. These primary public databases constitute the basis for accessibility, reproducibility, and reusability of genomic data. However, since quality assurance and correct assignment of taxonomy are the responsibility of data contributors, improving quality and taxonomic description has been an everlasting problem [3–5]. Low-quality data not only decrease 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 [6] 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 [7] estimates genome completeness and contamination by inspecting for the presence/absence of marker genes specific to each taxon. To confirm taxonomic affiliation of unidentified genomes, Bull *et al.* proposed using 16S rRNA genes together with housekeeping genes [4]. Beaz-Hidalgo *et al.* recommended the use of average nucleotide identity (ANI) to verify the taxonomic position of newly obtained

*Corresponding author. Masanori Arita, Center for Information Biology, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan. Phone: +81-55-981-9449; E-mail: arita@nig.ac.jp

genomes [8]. ANI represents the mean sequence identity of homologous regions between a given pair of genomes, and an ANI value of 95–96% is widely accepted as the threshold for distinguishing species [9–11]. Examples of ANI values and the 16S rRNA gene sequences for curated genomes can be available at the EzGenome and EzTaxon databases [12]. Recently, the use of genomic comparison methods including ANI was also proposed to find and correct misidentified genomes in the public databases at an NCBI workshop [13].

Along this line of research, we developed the DDBJ Fast Annotation and Submission Tool (DFAST) as a web-based bacterial annotation pipeline with integrated quality assessment using CheckM and taxonomic assessment using ANI. DFAST allows researchers to submit annotated genomes easily to INSDC through the DDBJ Mass Submission System (MSS) [14]. As the initial showcase of DFAST, we targeted lactic acid bacteria (LAB) and constructed a reference protein database tailored for *Lactobacillus* as well as *Pediococcus* to enable accurate and rapid annotation. We also developed an associated repository, DFAST Archive of Genome Annotation (DAGA), which stores LAB genomes obtained from DDBJ/ENA/GenBank and SRA with consistent annotation and assessment by DFAST. Our aim is to provide a reliable genome resource to the entire research community, thereby promoting accessibility and reusability of genomic data.

Among LAB, *Lactobacillus* is highly heterogeneous and the largest genus in the family *Lactobacillaceae*, comprising 185 species and 18 subspecies as of June 2016 (http://www.bacterio.net/lactobacillaceae.html). The genus *Pediococcus* is another member of *Lactobacillaceae* consisting of 11 species, and it is phylogenetically placed within the *Lactobacillus* cluster, near *L. plantarum* and *L. brevis* [15, 16]. In a recent study, the term *Lactobacillus sensu lato* was also proposed to refer to these genera [17]. In both genera, the number of new species described and genomes published have been growing with the improvement of isolation, cultivation, and identification methods as well as sequencing technology (Fig. 1). 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 Sanger Institute, UK (PRJEB3060), and "Genomic characterization of the genus *Lactobacillus*" in China (PRJNA222257). The results of such projects have enabled genus-wide analyses covering almost 90% of the known species based on the genomic information [17, 18]. Our DAGA provides annotated genomes for the family *Lactobacillaceae*, which includes many species that have undergone reclassification and species difficult to distinguish by 16S rRNA gene sequences. Our data will benefit all researchers who use LAB genomes, especially those focusing on inter- and intraspecific relations.

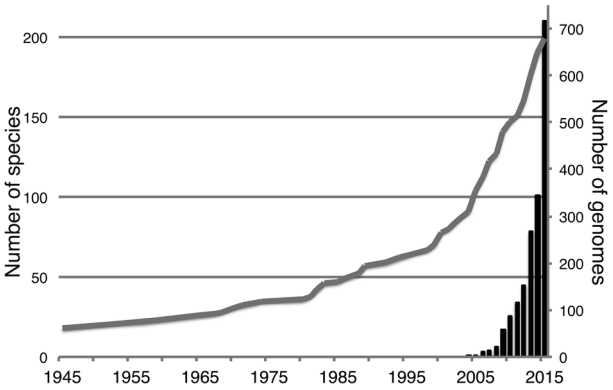

Fig. 1. The number of described species and published genomes in *Lactobacillus* and *Pediococcus*.

Solid line represents the cumulative number of described (sub) species. Only valid species as of Jan. 2016 were included, not reclassified ones. The bar chart represents the cumulative number of genomes deposited in DDBJ/ENA/GenBank.

In the present article, we describe development of DAGA and DFAST, and we also report several findings related to the current nomenclature.

MATERIALS AND METHODS

Construction of the annotation pipeline

The reference protein database was first constructed to provide consistent annotation to all focused genomes. A total of 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 other genomes were added to link with the *Lactobacillales*-specific Clusters of Orthologous Genes (LaCOGs) [19] and Microbial Genome Database (MBGD) [20]: *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

Fig. 2. Screenshots of DFAST and 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. D) Submission form of DFAST. Users can annotate their own genome by uploading the FASTA file. E) Result of DFAST. Submission files for DDBJ Mass Submission System are ready.

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 its default settings [21]. 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 [22]. 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. A total of 6,428 clusters were assigned to LaCOGs, of which 98.9% formed a one-to-one relationship with specific LaCOG clusters. Likewise, an additional 1,601 clusters were assigned to MBGD, of which 94.4% were one-to-one. To confirm the protein functions, public protein databases and the NCBI Conserved Domain Database [23] were searched manually. All protein names followed the NCBI guidelines for naming proteins (http://www.ncbi.nlm.nih.gov/genbank/genomesubmit_annotation/).

The core annotation process was based on the Prokka annotation software [24], performing prediction of tRNAs, rRNAs, CRISPRs, and protein-coding sequences as well as similarity searches against protein sequence databases and protein family profiles. The reference database was used in our customized Prokka pipeline that can generate DDBJ-compliant submission files.

Data collection

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 [25]. Raw sequence data (Illumina sequences with the paired-end method) were downloaded from SRA, and *de novo* assembly was conducted to reconstruct draft genome sequences as described below. All genomes were annotated with the customized Prokka pipeline.

Genome assembly

Raw sequence reads were preprocessed using Platanus_trim (version 1.0.7) to remove low-quality regions. *De novo* assembly was conducted using the Platanus assembler (version 1.2.4) [26]. Since Platanus was originally developed for heterozygous diploid genomes, we specified the parameters "-d 0.3 -u 0.05" to configure it 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.

Calculation of average nucleotide identity

The pyani script (https://github.com/widdowquinn/pyani) was used to calculate the ANI between two genomes, based on the method by Goris *et al.* [9]. In brief, one genome was cut into 1,020 nt fragments, which were searched against the other genome by using the BLASTN algorithm [27]. 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 Fig. 3 (B–D) were constructed by the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) clustering method with a distance of (1 – ANI).

Quality assessment of genomes

CheckM (version 1.0.5) was used to calculate completeness and contamination of each genome [7]. 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.

Implementation of the web service

DFAST and DAGA were implemented in Python 2.7.11 with PostgreSQL 8.4.20 and Nginx 1.8.0, and run on a Red Hat Enterprise Linux server (release 6.7).

RESULTS

Overview of the DAGA service

We developed an integrated genome archive specialized for LAB, namely DAGA. The first version of the dataset targets 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

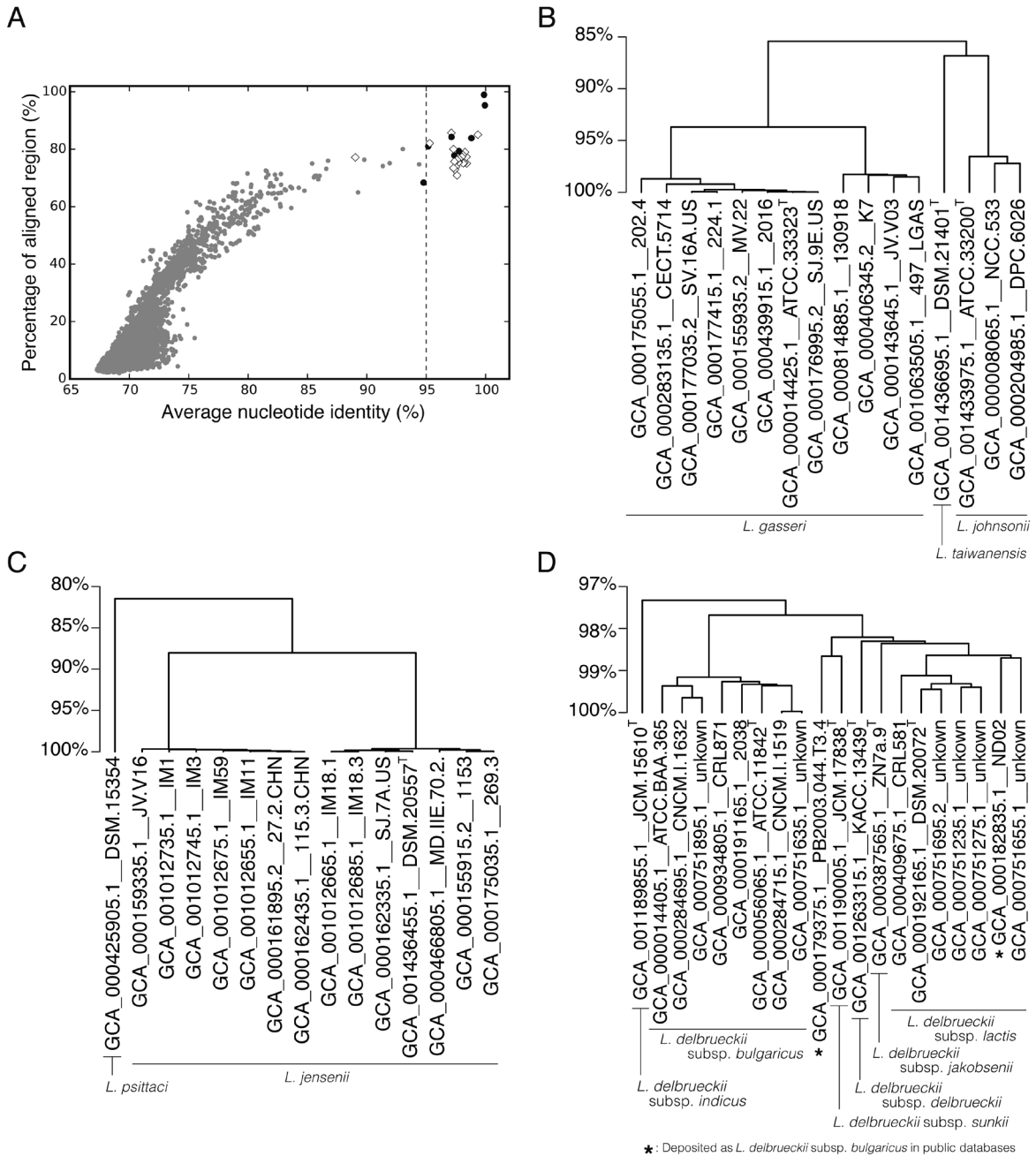

Fig. 3. A) Distribution of ANI values among 191 representative genomes. Circles and open diamonds indicate interspecific and intraspecific ANI values, respectively. Black circles indicate problematic genomes. B–D) Hierarchical clustering results by using (1 – ANI) as the genome distance. Each label represents the accession number of NCBI Assembly Database and the strain name. B: *Lactobacillus gasseri*, C: *L. jensenii*, D: *L. delbrueckii*.

de novo from raw reads deposited in SRA. All genomes were annotated by the Prokka (ver. 1.11) pipeline with the customized reference database for LAB. The quality of genomes was assessed by examining the presence of specific gene markers with CheckM, and taxonomic affiliation was verified by ANI. As of January 2016, DAGA covers 168 species and 18 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" in the genome identifier are from the NCBI Assembly Database, and those with "DRR", "ERR", or "SRR" are from SRA.

Figure 2 shows screenshots of DAGA. Users can query genomes of interest from the search form in the upper part or select a 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 for each grade are shown in Tables 1 and 2. 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.

Selection of a representative genome for each species

To verify the taxonomic relationship of each species, we calculated pairwise ANI values among 191 strains representing each species (or subspecies). We gave priority 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 3A shows the results of ANI calculation (also see our website https://dfast.nig.ac.jp/download). In most cases, the ANI values between species were below 95%, the threshold to differentiate species. Six strains in Table 3 (black circles in the Fig. 3A) showed anomalously high ANI values, indicating the incongruence of their taxonomic positions, which will be discussed later.

By excluding these six strains, we obtained 185 representative genomes whose interspecific pairwise ANI values were well below 95%. One exception was *L. zeae* and *L. casei*, which had an ANI of 94.4% (see Discussion). After a long period of controversy, *L. zeae* is now considered to be in the same taxon as *L. casei* [28]. However, the organism name has not been formally rejected in the current nomenclature, and *L. zeae* was counted with its original name in our database. It should also be noted that the publicly available genome of *L. amylotrophicus* (GCA_001434555.1), which exhibited an ANI of 99.9% with *L. amylophilus*, did not serve as the representative genome. Instead, we used data from SRA (ERR387486) as the representative of *L. amylotrophicus*.

The validity of the 185 representative genomes was also confirmed by comparing their reconstructed 16S rRNA gene sequences with those deposited in public databases. When not available, housekeeping genes like *pheS* or *rpoA* were used instead. In addition, a phylogenetic tree was constructed using 132 conserved single-copy genes to verify their taxonomic positions, and this tree is available at our website (https://dfast.nig.ac.jp/download/). Selection of representative genomes was implemented as a procedure in our system to serve as a tool for taxonomic studies in which comparison with type strains is critical.

Detection of mislabeled genomes by ANI values

We next checked the taxonomic affiliation for all genomes in DAGA by conducting ANI calculations against the representative genomes. We adopted species names based on the ANI calculations for 77 mislabeled genomes and inferred names for 55 unidentified genomes that were deposited as *Lactobacillus* sp. Such genomes with problematic taxonomic positions were marked as Rating 1 (Table 4).

Remarkably, 28 of 32 "*L. casei*" genomes were in fact *L. paracasei*, as previously postulated in the literature [29] 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 type strains, one was low quality with 22% ambiguous bases (N), and the last was the recently published *L. casei* N87 (GCA_001013375.1). The last strain shared an ANI of 96.8% with *L. zeae* DSM 20178^{T} and an ANI of 94.3% with *L. casei* ATCC 393 ^{T}. 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 inferred to be *L. plantarum* subsp. *argentoratensis*, *L. pentosus*, and *L. paraplantarum*, respectively. All assignments were recorded, i.e., both the original and the corrected names are available in our database.

Genomic diversity of LAB revealed by ANI

As a demonstrative analysis taking advantage of the wealth of genomic data stored in DAGA, we conducted all-against-all ANI comparison between 704 genomes (N = 704 × 703/2 = 247,456) to further investigate genomic diversity. Low-quality genomes and genomes with ambiguous taxonomy were excluded. All interspecific ANI values (N=239,840) were less than 95%, while 198 out of the remaining 7,616 intraspecific ANI values were also less than 95%. Such exceptions included the divergence within *L. kunkeei*, *L. gasseri*, and *L. jensenii*. *L. gasseri* and *L. jensenii* were each clearly separated into

Table 1. 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 2. 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

Table 3. Strains with problematic taxonomic positions

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

[#] Non-type strain.

Table 4. Mislabeled genomes deposited in DDBJ/ENA/GenBank

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

* Those with GCA were derived from NCBI Assembly Database and those with DRR/SRR/ERR were derived from SRA.

two previously unknown subgroups (Figs. 3B and 3C). The ANI values between the subgroups were 93% and 88% for *L. gasseri* and *L. jensenii*, respectively, while the ANI values within the same subgroups were over 98% in both species. The intraspecific separation was also supported by the multiple alignments of their *pheS* gene sequences (alignment data not shown). For *L. gasseri* and *L. jensenii*, the nucleotide identities of *pheS* genes between the subgroups were 96% and 93%, while those of *rpoA* genes were 99% and 98%, respectively. The intraspecific separation in the two species might deserve subspecies-level differentiation. We must note, however, that our analysis was based on genomic information only. Further analysis including phenotypic characterization is required to establish their valid classifications.

To assess the discriminating power of ANI, ANI values were calculated among six subspecies of *L. delbrueckii*. The ANI values for their type strains were distributed in the range of 97.2–98.4%. In spite of such high values, hierarchical clustering based on the ANI values could separate them (Fig. 3D), and the tree topology was roughly consistent with the ones from multilocus sequence analyses [30, 31]. This implies the reliability of ANI in evaluating the genetic subgroups within a species.

DFAST online annotation server

We developed a web interface for the DFAST annotation pipeline, so that users can manage metadata and submit annotated genomes to DDBJ. Users can annotate their own genomic data by uploading a FASTA formatted file via a submission form and can perform quality and taxonomic assessment using CheckM and the calculation of ANI. A simple annotation editor is also available, allowing users to modify gene product names or gene symbols. Submission files for the DDBJ Mass Submission System are then automatically generated. Results can be downloaded in several formats, including GenBank, Multi-FASTA, and tab-separated formats.

DISCUSSION

Recent new-generation sequencing technologies are constantly producing more and more genome sequences, making it important to assess their data quality and taxonomic positions. DAGA is a new genome archive that stores quality-controlled and taxonomically confirmed bacterial genomes with consistent annotation. Its quality measure is the genome completeness and contamination values calculated by CheckM, and we were able to use it to successfully identify genomes of incorrect size as compared with typical LAB strains without using any other selection method. In addition, we also identified taxonomically mislabeled genomes in public databases even for type strains (Table 2A). These results will help researchers to select genomes for comparative analysis.

The NCBI Reference Sequence (RefSeq) and the Pathosystems Resource Integration Center (PATRIC) provide consistently annotated genome collections [32, 33]. They collect genome sequences from DDBJ/ENA/GenBank and re-annotate them using NCBI Prokaryotic Genome Annotation Pipeline (PGAP) 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. Because SRA stores raw sequence data, it is difficult for users without bioinformatics skills to exploit the data. DAGA facilitates the reuse of valuable data available in SRA, such as the only reliable genome for *L. amylotrophicus*, which can only be obtained from SRA (ERR387486).

As of January 2016, DAGA provides 1,421 genomes collected from DDBJ/ENA/GenBank and SRA for two genera in *Lactobacillaceae*. The genus *Sharpea* was not included even though it is 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* [34, 35]. As the number of available genomes is increasing rapidly, we plan to update the database regularly and to expand the scope of the database to other taxonomic groups.

The most widely used methodology for bacterial taxonomic identification is the combination of 16S rRNA gene sequencing and DNA-DNA hybridization (DDH) [36]. According to the minimal standard recommended for describing new species of *Lactobacillus*, DDH should be conducted if the 16S rRNA sequence similarity to the closest known species is beyond 97% [37]. Recently, however, ANI has been used as a substitute for DDH to describe novel species of *Lactobacillus* [38–40]. ANI has several advantages. First, it does not require a laboratory assay and is computationally reproducible. Second, it 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 difficult to reconstruct. Lastly and most importantly, ANI shows prominent discriminatory power to determine genome identity. Even between hard-to-distinguish taxonomic groups such as *L. casei* and *L. plantarum*, the ANI values between two different species

were below 85%, much less than the threshold of 95%. Furthermore, only 0.4% of the comparisons fell within the "twilight zone" of 85–95% in our analysis of 191 representative genomes (Fig. 3A). For these reasons, we emphasize the benefit of ANI to validate taxonomic status for genomes deposited in DAGA. As an exception, the ANI value between *L. casei* ATCC 393^T and *L. zeae* DSM 20178^T was slightly below the species-level threshold (94.4%) even though the two strains are now considered the same species. In our analysis, ANI values between species were always less than 95%, but the reverse is not always true. As shown by the results for *L. gasseri* and *L. jensenii,* intraspecific ANI values can be lower than 95% in some species.

In several species, ANI can help determine subspecies of a given strain, as shown in the results for *L. delbrueckii* (Fig. 3D). It seems difficult to establish an ANI cutoff value to distinguish subspecies, however, because inter-subspecific ANI values depend on the species (Fig. 3A). For example, the lowest value exhibited by the subspecies of *L. aviarius* was 89%, much lower than the species-level threshold. The highest value was reported for *L. kefiranofaciens*, which showed an ANI value as high as 99.4%. According to the original description, the two subspecies of *L. kefiranofaciens* shared 100% 16S rRNA sequence identity and were distinguishable by morphological and biochemical characteristics [41]. On the other hand, the two subspecies of *L. plantarum* were distinguished mainly based on their genotypic traits because their morphological, physiological, and biochemical characteristics were almost identical, with the only exceptions being in a few carbohydrate fermentation patterns [42]. The ANI value between *L. plantarum* subsp. *plantarum* and subsp. *argentoratensis* was 95.3%. The difference in inter-subspecific ANI values between the two species seems to reflect their original descriptions. For several strains, the allocation of subspecies was found to be inconsistent with the ANI results. *L. sakei* subsp. *sakei* 23 K was more similar to subsp. *carnosus* than to subsp. *sakei*, as suggested by Chaillou *et al.* [43]. The two strains labeled as *L. paracasei* subsp. *tolerans* (GCA_000409835.1 and GCA_000410335.1) were more similar to subsp. *paracasei*, although the difference was as small as 0.2%. The genome sizes of subsp. *paracasei* and subsp. *tolerans* differ prominently: 3.0 Mbp and 2.4 Mbp, respectively. Judging from the genome sizes, the two strains are more likely to belong to the subspecies *paracasei*. However, we could not find any other evidence that supports this assumption. The values from all ANI calculations are available from our website: https://dfast.nig.ac.jp/download/.

Our assessment found the six questionable genomes listed in Table 2A, 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). The *P. lolii* genome was presumably a misclassification of the sequenced strain. A previous study reported that the type strains of *P. lolii* deposited in DSMZ and JCM were strains of *Pediococcus acidilactici* [44]. Our 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 [45]. This species is displayed as *Lactobacillus* sp. and *Lactobacillus hilgardii* in JCM and the EzGenome database, respectively [12, 46]. The type strains of *L. homohiochii* and *L. fornicalis* deposited in culture collections were reported to misrepresent the originally described strains [47] (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.* Our analysis suggests that its large genome size [18] and the similarity to *L. kefiri* [17] are attributable to the sequence contamination. Likewise, the genome of *P. parvulus* DSM 20332^T seems to be contaminated with another strain of *P. acidilactici*.

Our annotation pipeline is freely available as the DFAST web service. In comparison with other annotation tools such as RAST [48] or the Microbial Genome Annotation Pipeline (MiGAP) [49], the advantage of DFAST is the ability to generate ready-to-submit annotation files. RAST can perform detailed functional annotation based on the platform called SEED. However, if users want to submit an annotated genome to INSDC, they need to convert annotation results into an acceptable format. Although MiGAP partly supports the DDBJ-acceptable format, users are still required to prepare metadata and to curate annotated protein names before submission. As our curated reference database follows the protein naming guidelines of the NCBI, minimal manual curation, if any,

is required before submitting genomes to DDBJ. Another advantage of DFAST is its short running time. It takes about 5 minutes to annotate a typical 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. We have already used DFAST to annotate 5 genomes of *Lactobacillus* strains, including two candidates for new species (manuscript in preparation). On average, 90.3% of protein coding sequences were annotated based on a similarity search against the reference protein database in this study. We were able to submit them to DDBJ without any manual curation. Currently, the reference database constructed in this study is based mainly on protein sequence data obtained from *Lactobacillus* and *Pediococcus*, with additional information from 12 representative strains of other genera. Our future tasks include an update and extension of the reference database to other genera, such as *Lactococcus* and *Leuconostoc*, and annotation of frameshifted genes or pseudo-genes.

In conclusion, we assessed 1,421 genomes covering 191 (sub)species in the family *Lactobacillaceae* and developed a curated genome repository referred to as DAGA. This will improve the accessibility and reusability of LAB genome resources. The annotation and submission pipeline DFAST will help researchers to deal with large amounts of emerging sequence data, thereby accelerating studies of LAB that make use of genomic data.

ACKNOWLEDGEMENTS

We gratefully acknowledge Masanori Tohno and Akihito Endo for helpful discussion and informative suggestions. We also thank Kyungbum Lee and Toshihisa Okido at DDBJ Center for helpful comments. The pyani script was kindly provided by Leighton Pritchard at James Hutton Institute, UK. This work was supported by Collaborative Research Program A (2014–2015) of the National Institute of Genetics (NIG) and the commission for Development of Artificial Gene Synthesis Technology for Creating Innovative Biomaterial of the Ministry of Economy, Trade and Industry, Japan. Computational analysis was performed on the NIG supercomputer at the Research Organization of Information and Systems (ROIS).

REFERENCES

1. Cochrane G, Karsch-Mizrachi I, Takagi T, International Nucleotide Sequence Database Collaboration. 2016. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res 44 D1: D48–D50. [Medline] [CrossRef]
2. Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database Collaboration. 2012. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res 40: D54–D56. [Medline] [CrossRef]
3. Nilsson RH, Ryberg M, Kristiansson E, Abarenkov K, Larsson KH, Kõljalg U. 2006. Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. PLoS One 1: e59. [Medline] [CrossRef]
4. Bull MJ, Marchesi JR, Vandamme P, Plummer S, Mahenthiralingam E. 2012. Minimum taxonomic criteria for bacterial genome sequence depositions and announcements. J Microbiol Methods 89: 18–21. [Medline] [CrossRef]
5. Nakazato T, Ohta T, Bono H. 2013. Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive. PLoS One 8: e77910. [Medline] [CrossRef]
6. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29: 1072–1075. [Medline] [CrossRef]
7. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25: 1043–1055. [Medline] [CrossRef]
8. Beaz-Hidalgo R, Hossain MJ, Liles MR, Figueras MJ. 2015. 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. [Medline] [CrossRef]
9. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 57: 81–91. [Medline] [CrossRef]
10. Richter M, Rosselló-Móra R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci USA 106: 19126–19131. [Medline] [CrossRef]
11. Kim M, Oh HS, Park SC, Chun J. 2014. 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. [Medline] [CrossRef]
12. Kim OS, Cho YJ, Lee K, Yoon SH, Kim M, Na H, Park SC, Jeon YS, Lee JH, Yi H, Won S, Chun J. 2012. Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. Int J Syst Evol Microbiol 62: 716–721. [Medline] [CrossRef]
13. Federhen S, Rosselló-Mora R, Klenk HP, Tindall BJ, Konstantinidis KT, Whitman WB, Brown D, Labeda

D, Ussery D, Garrity GM, Rita R, Colwell NH, Graf J, Parte A, Yarza P, Goldberg B, Sichtig H, Karsch-Mizrachi I, Clark K, McVeigh R, Pruitt KD, Tatusova T, Falk R, Turner S, Madden T, Kitts P, Kimchi A, Klimke W, Agarwala R, DiCuccio M, Ostell J. 2016. Meeting report: GenBank microbial genomic taxonomy workshop (12–13 May, 2015). Stand Genomic Sci 11: 15. [CrossRef]

14. Sugawara H, Miyazaki S, Gojobori T, Tateno Y. 1999. DNA Data Bank of Japan dealing with large-scale data submission. Nucleic Acids Res 27: 25–28. [Medline] [CrossRef]
15. Zhang ZG, Ye ZQ, Yu L, Shi P. 2011. Phylogenomic reconstruction of lactic acid bacteria: an update. BMC Evol Biol 11: 1. [Medline] [CrossRef]
16. Franz CMAP, Endo A, Abriouel H, Van Reenen CA, Gálvez A, Dicks LMT. 2014. The genus *Pediococcus*, pp. 359–376. *In* Holzapfel, WH, Wood, BJB (eds.), Lactic Acid Bacteria: Biodiversity and Taxonomy. John Wiley & Sons.
17. Zheng J, Ruan L, Sun M, Gänzle M. 2015. A genomic view of lactobacilli and pediococci demonstrates that phylogeny matches ecology and physiology. Appl Environ Microbiol 81: 7233–7243. [Medline] [CrossRef]
18. Sun Z, Harris HMB, McCann A, Guo C, Argimón S, Zhang W, Yang X, Jeffery IB, Cooney JC, Kagawa TF, Liu W, Song Y, Salvetti E, Wrobel A, Rasinkangas P, Parkhill J, Rea MC, O'Sullivan O, Ritari J, Douillard FP, Paul Ross R, Yang R, Briner AE, Felis GE, de Vos WM, Barrangou R, Klaenhammer TR, Caufield PW, Cui Y, Zhang H, O'Toole PW. 2015. Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. Nat Commun 6: 8322. [Medline] [CrossRef]
19. Makarova K, Slesarev A, Wolf Y, Sorokin A, Mirkin B, Koonin E, Pavlov A, Pavlova N, Karamychev V, Polouchine N, Shakhova V, Grigoriev I, Lou Y, Rohksar D, Lucas S, Huang K, Goodstein DM, Hawkins T, Plengvidhya V, Welker D, Hughes J, Goh Y, Benson A, Baldwin K, Lee JH, Díaz-Muñiz I, Dosti B, Smeianov V, Wechter W, Barabote R, Lorca G, Altermann E, Barrangou R, Ganesan B, Xie Y, Rawsthorne H, Tamir D, Parker C, Breidt F, Broadbent J, Hutkins R, O'Sullivan D, Steele J, Unlu G, Saier M, Klaenhammer T, Richardson P, Kozyavkin S, Weimer B, Mills D. 2006. Comparative genomics of the lactic acid bacteria. Proc Natl Acad Sci USA 103: 15611–15616. [Medline] [CrossRef]
20. Uchiyama I. 2007. MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. Nucleic Acids Res 35: D343–D346. [Medline] [CrossRef]
21. Contreras-Moreira B, Vinuesa P. 2013. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol 79: 7696–7701. [Medline] [CrossRef]
22. Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178–2189. [Medline] [CrossRef]
23. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH. 2015. CDD: NCBI's conserved domain database. Nucleic Acids Res 43: D222–D226. [Medline] [CrossRef]
24. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30: 2068–2069. [Medline] [CrossRef]
25. Kitts PA, Church DM, Thibaud-Nissen F, Choi J, Hem V, Sapojnikov V, Smith RG, Tatusova T, Xiang C, Zherikov A, DiCuccio M, Murphy TD, Pruitt KD, Kimchi A. 2016. Assembly: a resource for assembled genomes at NCBI. Nucleic Acids Res 44 D1: D73–D80. [Medline] [CrossRef]
26. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, Kohara Y, Fujiyama A, Hayashi T, Itoh T. 2014. Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res 24: 1384–1395. [Medline] [CrossRef]
27. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215: 403–410. [Medline] [CrossRef]
28. Judicial Commission of the International Committee on Systematics of Bacteria. 2008. 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. [Medline] [CrossRef]
29. Smokvina T, Wels M, Polka J, Chervaux C, Brisse S, Boekhorst J, van Hylckama Vlieg JET, Siezen RJ. 2013. *Lactobacillus paracasei* comparative genomics: towards species pan-genome definition and exploitation of diversity. PLoS One 8: e68731. [Medline] [CrossRef]
30. Tanigawa K, Watanabe K. 2011. Multilocus sequence typing reveals a novel subspeciation of *Lactobacillus delbrueckii*. Microbiology 157: 727–738. [Medline] [CrossRef]
31. Adimpong DB, Nielsen DS, Sørensen KI, Vogensen FK, Sawadogo-Lingani H, Derkx PMF, Jespersen L. 2013. *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. [Medline] [CrossRef]
32. Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I. 2015. RefSeq microbial genomes database: new representation and annotation strategy. Nucleic Acids Res 43: 3872. [Medline] [CrossRef]
33. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix D, Kenyon R, Machi D, Mao C, Nordberg EK, Olson R, Overbeek R, Pusch GD, Shukla M, Schulman J, Stevens RL, Sullivan DE, Vonstein V, Warren A, Will R, Wilson MJC, Yoo HS, Zhang C, Zhang Y, Sobral BW. 2014. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res 42: D581–D591. [Medline] [CrossRef]
34. Morita H, Shiratori C, Murakami M, Takami H, Toh H, Kato Y, Nakajima F, Takagi M, Akita H, Masaoka T, Hattori M. 2008. *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. [Medline] [CrossRef]
35. Salvetti E, Felis GE, Dellaglio F, Castioni A, Torriani S, Lawson PA. 2011. 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. [Medline] [CrossRef]
36. Chun J, Rainey FA. 2014. Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. Int J Syst Evol Microbiol 64: 316–324. [Medline] [CrossRef]
37. Mattarelli P, Holzapfel W, Franz CMAP, Endo A, Felis GE, Hammes W, Pot B, Dicks L, Dellaglio F. 2014. Recommended minimal standards for description of new taxa of the genera *Bifidobacterium, Lactobacillus* and related genera. Int J Syst Evol Microbiol 64: 1434–1451. [Medline] [CrossRef]
38. Olofsson TC, Alsterfjord M, Nilson B, Butler E, Vásquez A. 2014. *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. [Medline] [CrossRef]
39. Mao Y, Chen M, Horvath P. 2015. *Lactobacillus herbarum* sp. nov., a species related to *Lactobacillus plantarum*. Int J Syst Evol Microbiol 65: 4682–4688. [Medline] [CrossRef]
40. Puertas AI, Arahal DR, Ibarburu I, Elizaquível P, Aznar R, Dueñas MT. 2014. *Lactobacillus sicerae* sp. nov., a lactic acid bacterium isolated from Spanish natural cider. Int J Syst Evol Microbiol 64: 2949–2955. [Medline] [CrossRef]
41. Vancanneyt M, Mengaud J, Cleenwerck I, Vanhonacker K, Hoste B, Dawyndt P, Degivry MC, Ringuet D, Janssens D, Swings J. 2004. Reclassification of *Lactobacillus kefirgranum* Takizawa et al. 1994 as *Lactobacillus kefiranofaciens* subsp. *kefirgranum* subsp. nov. and emended description of *L. kefiranofaciens* Fujisawa et al. 1988. Int J Syst Evol Microbiol 54: 551–556. [Medline] [CrossRef]
42. Bringel F, Castioni A, Olukoya DK, Felis GE, Torriani S, Dellaglio F. 2005. *Lactobacillus plantarum* subsp. *argentoratensis* subsp. nov., isolated from vegetable matrices. Int J Syst Evol Microbiol 55: 1629–1634. [Medline] [CrossRef]
43. Chaillou S, Daty M, Baraige F, Dudez AM, Anglade P, Jones R, Alpert CA, Champomier-Vergès MC, Zagorec M. 2009. Intraspecies genomic diversity and natural population structure of the meat-borne lactic acid bacterium *Lactobacillus sakei*. Appl Environ Microbiol 75: 970–980. [Medline] [CrossRef]
44. Wieme A, Cleenwerck I, Van Landschoot A, Vandamme P. 2012. *Pediococcus lolii* DSM 19927^T and JCM 15055^T are strains of *Pediococcus acidilactici*. Int J Syst Evol Microbiol 62: 3105–3108. [Medline] [CrossRef]
45. Skerman V, McGowan V, Sneath P. 1980. Approved lists of bacterial names. Int J Syst Bacteriol 30: 225–420. [CrossRef]
46. Kitahara M. 2008. Quality management of *Lactobacillus* strains in JCM. Microbiol Cult Collect 24: 143–145.
47. Goto N, Joyeux A, Lonvaud-Funel A. 1994. Taxonomic problem of the type strain of *Lactobacillus homohiochii*. J Brew Soc Jpn 89: 643–646. [CrossRef]
48. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R. 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res 42: D206–D214. [Medline] [CrossRef]
49. Sugawara H, Ohyama A, Mori H, Kurokawa K. 2009. Microbial genome annotation pipeline (MiGAP) for diverse users. *In* Proceedings of the 20th International Conference on Genome Informatics, Yokohama, pp. S-001-1-2.