



Comparative genome analysis of *Clostridium beijerinckii* strains isolated from pit mud of Chinese strong flavor baijiu ecosystem

Wei Zou ^{1,*}, Guangbin Ye,¹ Chaojie Liu,¹ Kaizheng Zhang,¹ Hehe Li,^{2,*} and Jiangan Yang ¹

¹College of Bioengineering, Sichuan University of Science & Engineering, Yibin, Sichuan 644005, China and

²Beijing Key Laboratory of Flavor Chemistry, Beijing Technology and Business University (BTBU), Beijing 100048, China

*Corresponding author: College of Bioengineering, Sichuan University of Science & Engineering, 1 Baita Road, Sanjiang New District, Yibin, Sichuan 644005, China. Emails: weizou1985@163.com (W.Z.); xyzhehe@126.com (H.L.)

Abstract

Clostridium beijerinckii is a well-known anaerobic solventogenic bacterium which inhabits a wide range of different niches. Previously, we isolated five butyrate-producing *C. beijerinckii* strains from pit mud (PM) of strong-flavor baijiu (SFB) ecosystems. Genome annotation of the five strains showed that they could assimilate various carbon sources as well as ammonium to produce acetate, butyrate, lactate, hydrogen, and esters but did not produce the undesirable flavors isopropanol and acetone, making them useful for further exploration in SFB production. Our analysis of the genomes of an additional 233 *C. beijerinckii* strains revealed an open pangenome based on current sampling and will likely change with additional genomes. The core genome, accessory genome, and strain-specific genes comprised 1567, 8851, and 2154 genes, respectively. A total of 298 genes were found only in the five *C. beijerinckii* strains from PM, among which only 77 genes were assigned to Clusters of Orthologous Genes categories. In addition, 15 transposase and 12 phage integrase families were found in all five *C. beijerinckii* strains from PM. Between 18 and 21 genome islands were predicted for the five *C. beijerinckii* genomes. The existence of a large number of mobile genetic elements indicated that the genomes of the five *C. beijerinckii* strains evolved with the loss or insertion of DNA fragments in the PM of SFB ecosystems. This study presents a genomic framework of *C. beijerinckii* strains from PM that could be used for genetic diversification studies and further exploration of these strains.

Keywords: baijiu; butyrate; *Clostridium beijerinckii*; mobile genetic elements; pangenome; pit mud

Introduction

Clostridium beijerinckii is a Gram-positive anaerobic solventogenic bacterium that is well known for its ability to produce acetone, butanol, and ethanol (ABE) or isopropanol, butanol, and ethanol (IBE) (Qureshi and Blaschek 2001; Survase et al. 2011; dos Santos Vieira et al. 2020) using low-cost carbon sources, such as different straw hydrolysates (Bellido et al. 2014; Dalal et al. 2019), lignocellulosic hydrolysate (Reddy et al. 2020), and molasses (Li et al. 2013; Fonseca et al. 2020). In addition, it has been used to produce hydrogen with great success (Seelert et al. 2015). Other metabolites that can be biosynthesized by wild or engineered *C. beijerinckii* strains include butyric acid (Drahokoupil and Patakova 2020), polyhydroxyalkanoate (Hassan et al. 2019), butyl butyrate, butyl acetate (Fang et al. 2020), and 1,3-propanediol (Wischnal et al. 2016). Furthermore, *C. beijerinckii* is widely used in co-cultures with other microorganisms for the production of butanol or hydrogen (Du et al. 2020).

Chinese baijiu, previously also called Chinese liquor or Chinese spirits, is a traditional distilled liquor which has existed in China for hundreds of years (Zheng and Han 2016; Jin et al. 2017; Liu and Sun 2018). Strong-flavor baijiu (SFB) is the most common type, accounting for more than 70% of total baijiu yields

(Xu et al. 2010). At present, more than 1300 flavor compounds have been identified in SFB (Yao et al. 2015), among which ethyl hexanoate, ethyl lactate, ethyl acetate, and ethyl butanoate are the representative compounds (Zheng and Han 2016). Butanoate is an important precursor of ethyl butanoate (Xu et al. 2020), the content of which is of great importance to the quality of SFB. Many butyrate-producing strains have been isolated and identified in the pit mud (PM) of SFB ecosystems, and among them, *Clostridium* strains are considered the main producers (Zou et al. 2018a; Chai et al. 2019; Luo et al. 2019). Recently, we isolated five *C. beijerinckii* strains from the PM of different SFB plants in Sichuan province in China and found that they were capable of producing butyrate with yields of 3.7–6.8 gL⁻¹ (Tian et al. 2019). These *C. beijerinckii* strains are candidate baijiu additives for butyrate production and potentially for cultivating new PM. Moreover, they could be of further use in the industrial fermentation of butyrate, butanol, and hydrogen. However, due to the uniqueness of the PM habitat, the physiological and metabolic features of the five *C. beijerinckii* strains may be different from other *C. beijerinckii* isolates, which may hinder their further exploration.

The huge amount of genomic information and the number of available bioinformatics software and databases allows for the

Received: June 07, 2021. Accepted: August 26, 2021

© The Author(s) 2021. Published by Oxford University Press on behalf of Genetics Society of America.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (<https://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

analysis of genomes across species or genera (Zhong et al. 2017; Jiang et al. 2020). In addition, pangenomic methods can be carried out for the analysis of the genome of a species or genus or a given phylogenetic clade (Udaondo et al. 2017). Pangenomes can contribute to our understanding of species diversity and the metabolic capabilities of species isolated from different niches (Vernikos et al. 2015; Golicz et al. 2020).

In this study, we sequenced four *C. beijerinckii* strains isolated from PM, named 3-8, G3-1, G3-3, and G3-5 (Tian et al. 2019). The four genomes as well as the previously sequenced genome of strain 2-1 were analyzed and compared. The metabolic features of the five strains related to assimilation of different carbon sources and the biosynthetic pathways of butyrate, butanol, and hydrogen were studied. Then, a pangenomic analysis of *C. beijerinckii* species was carried out to analyze the phylogenetic relationships, unique gene functions, and mobile genetic elements (MGEs) of the five *C. beijerinckii* strains isolated from PM.

Materials and methods

DNA extraction, genome sequencing, assembly, and annotation

Four *C. beijerinckii* strains, 3-8, G3-1, G3-3, and G3-5, were isolated from the PM of Chinese baijiu ecosystems (Tian et al. 2019). The four strains were grown on reinforced Clostridium medium (Munch-Petersen and Boundy 1963) and cultivated at 35°C for 60 h. Bacterial cells were then collected by centrifugation at 8000 rpm for 10 min, and genomic DNA was extracted using a previously described method (Tian et al. 2019).

The entire genomes of all four *C. beijerinckii* strains were sequenced using an Illumina NovaSeq platform by Shanghai Personal Biotechnology Co., Ltd. TruSeq™ DNA Sample Prep Kit was used to prepare the Illumina libraries. After obtaining the raw sequence data, AdapterRemoval (version 2.1.7) (Schubert et al. 2016) and SOAPec (version 2.0) (Luo et al. 2012) were used for data filtering and quality adjustments. Quality sequenced reads were then assembled using A5-MiSeq (Coil et al. 2015) and SPAdes (Bankevich et al. 2012), and base correction were performed on the assembled data using Pilon (Walker et al. 2014) to obtain the final assemblies. Gene prediction and annotation were carried out for the four *C. beijerinckii* genomes sequenced in this study and the previously sequenced genome of strain 2-1 using the Rapid Annotation using Subsystem Technology (RAST) server (Aziz et al. 2008). The amino acid sequences of proteins were further annotated using the Kyoto Encyclopedia of Genes and Genome (KEGG) Automatic Annotation Server (KAAS) pipeline (Moriya et al. 2007). Metabolic pathways of *C. beijerinckii* were constructed using KEGG Mapper and created using the KAAS pipeline. MGEs, including genome islands (GIs) and prophage sequences, were predicted for the five *C. beijerinckii* genomes. GIs were predicted using IslandViewer 4, which involves three methods: SIGI-HMM, IslandPath-DIMOB, and IslandPick (Bertelli et al. 2017). Prophage sequences were predicted using PHASTER (Arndt et al. 2016). Default parameters were used for all software in this study unless otherwise specified.

Pangenome analysis

The genomes of an additional 233 *C. beijerinckii* strains were used for pangenomic analysis, including 229 genomes download from the National Center for Biotechnology Information (NCBI) and the genomes of the four *C. beijerinckii* strains sequenced in this study. The detailed genomic information used in this study is provided in Supplementary Table S1. Pangenome analysis of the

233 *C. beijerinckii* strains was performed using the Bacterial Pan Genome Analysis (BPGA) pipeline (version 1.3) (Chaudhari et al. 2016). USEARCH was used as the clustering tool, and the sequence identity cutoff was set to 50%. A gene presence-absence binary matrix (pan-matrix) was obtained from BPGA and input into PanGP to analyze the pangenome and core genome profiles (Zhao et al. 2014). The mathematical formula used for pangenome profile fitting is a power-law regression based on Heaps' law ($y = Ax^B + C$, where y represents the pangenome size; x represents the genome number; and A , B , and C are fitting parameters). When $0 < B < 1$, the pangenome size increases when new genomes are analyzed, and the pangenome is considered open. When $B > 1$, the pangenome size does not increase when new genomes are added and are considered closed. The mathematical formula used for fitting the core genome size is an exponential regression model ($y = Ae^{Bx} + C$, where y represents the core genome size; x represents the number of analyzed genomes; and A , B , and C are fitting parameters). eggNOG-mapper was used to annotate genes in the pangenome with Clusters of Orthologous Groups (COG) of proteins categories (Huerta-Cepas et al. 2019).

Phylogenetic tree reconstruction

To construct the *C. beijerinckii* phylogenetic tree based on the core genome, the amino acid sequences of the core genome were concatenated and aligned using the MAFFT online service (Katoh et al. 2019). A phylogenetic tree was constructed using PhyML with the maximum likelihood algorithm (Guindon et al. 2010). The phylogenetic tree was rendered using FigTree (version 1.4.4).

Results and discussion

Genome properties of *Clostridium beijerinckii* strains isolated from PM

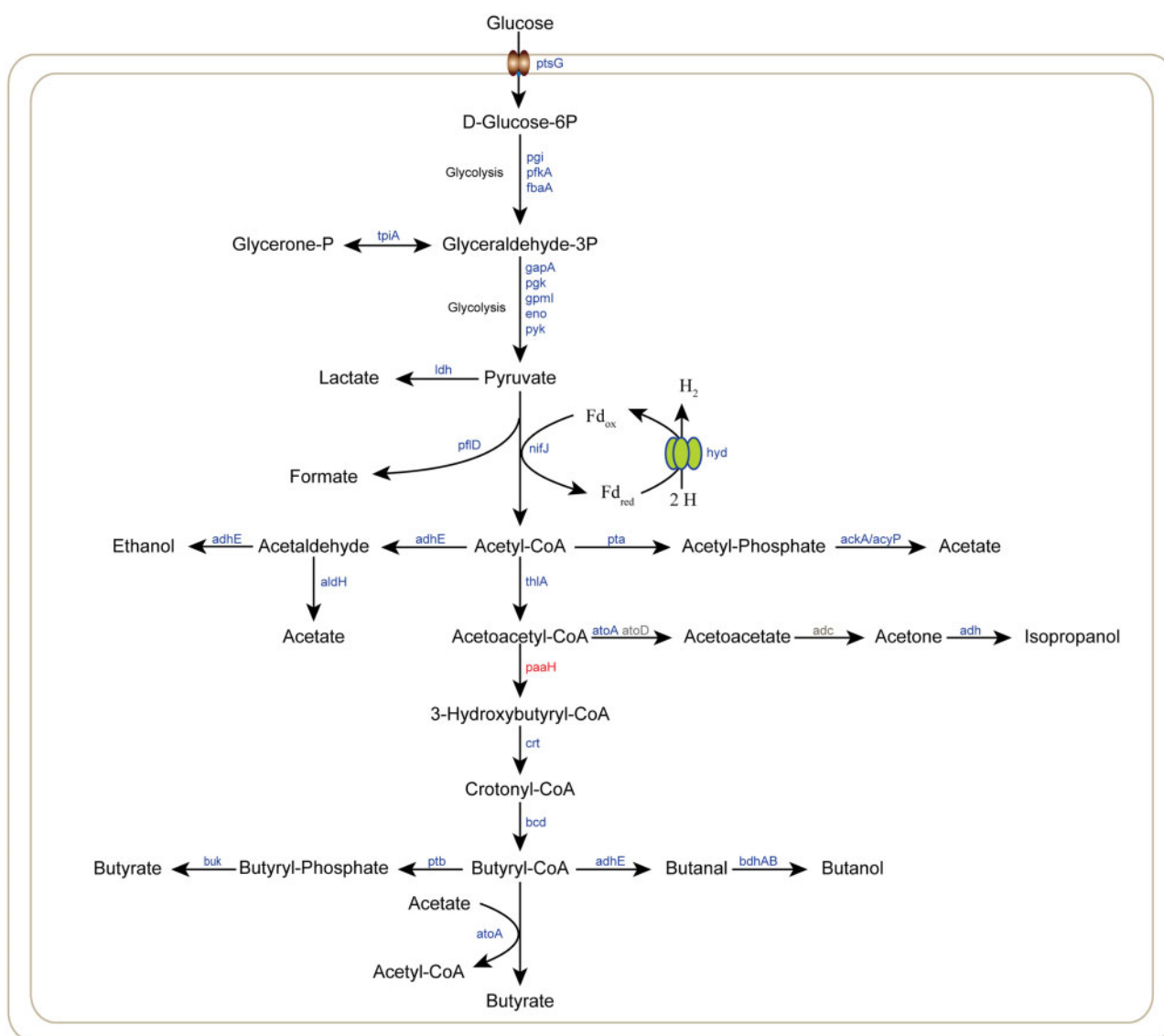
Previously, the genome sequence of *C. beijerinckii* strain 2-1 was sequenced, and the data were deposited into NCBI under accession number PRJNA428897. In this study, the genomes of *C. beijerinckii* strains 3-8, G3-1, G3-3, and G3-5 were sequenced and deposited into GenBank under accession numbers PRJNA690962, PRJNA695099, PRJNA695100, and PRJNA698586, respectively. The genome sizes of the five *C. beijerinckii* strains isolated from PM of SFB ecosystems ranged from 5,461,616 bp to 5,637,825 bp, and the number of scaffolds ranged from 92 to 328 (Table 1). The GC content of the five strains ranged from 29.58% to 29.78%. The RAST server was used to predict and annotate the genomes of the five isolated *C. beijerinckii* strains. The average number of annotated protein-encoding genes among the five genomes was 5040, and the average number of proteins annotated by RAST was 2039. The gene distributions according to RAST categories of the five *C. beijerinckii* strains were also compared (Supplementary Table S2). Carbohydrates, amino acids and derivatives, and cofactors, vitamins, prosthetic groups, and pigments were the three largest subsystems, representing an average of 17.0%, 12.4%, and 10.7% of each genome, respectively; Proteins belonging to the categories protein metabolism and dormancy and sporulation were less in strain 3-8 compared with other 4 strains (Supplementary Table S2).

Metabolic features of the five *Clostridium beijerinckii* strains from PM

The microbiota of PM in SFB ecosystems is an important inoculum for SFB production (Zou et al. 2018b). A variety of substrates are present during the saccharification and fermentation of cereals, mostly starch or hydrolysis products of starch such as disaccharides and monosaccharides. The main nitrogen source is

Table 1 Genome features of the five *Clostridium beijerinckii* strains isolated from pit mud (PM)

Features	2-1	3-8	G3-1	G3-3	G3-5
Genome size (bp)	5,626,308	5,461,616	5,637,955	5,609,807	5,637,825
No. of all scaffolds	328	92	129	105	130
Total reads	10,554,482	6,891,264	10,865,966	9,147,036	8,625,822
Total reads length (bp)	1,704,906,022	1,021,795,131	1,615,799,558	1,352,042,586	1,282,836,405
Largest scaffold length (bp)	246,790	235,144	235,510	279,509	235,510
Scaffold N50 (bp)	60,705	83,607	92,783	122,976	92,771
G+C content (%)	29.78	29.58	29.64	29.60	29.64
Coding protein number	5035	4932	5081	5071	5083
Proteins annotated belong RAST subsystems	2078	1869	2083	2084	2083
rRNA	6	8	37	17	36
tRNA	59	52	81	58	80

**Figure 1** Biosynthetic pathways of acetone, butanol, and ethanol (ABE) and isopropanol, butanol, and ethanol (IBE) from glucose in the five *Clostridium beijerinckii* strains isolated from pit mud (PM). Genes shown in gray were absent in all five strains. Genes shown in red were absent in strain 3-8 but present in strains 2-1, G3-1, G3-3, and G3-5.

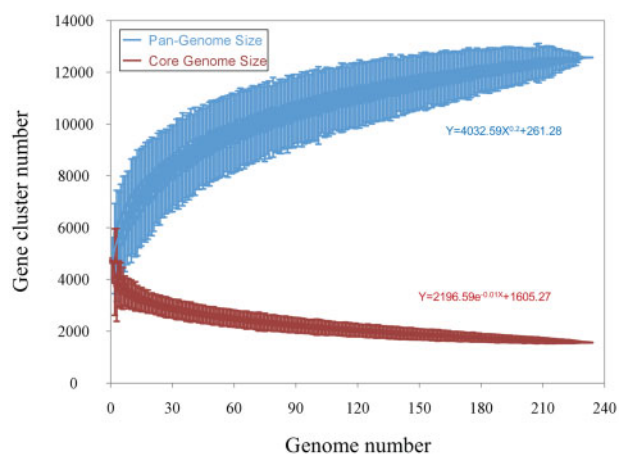


Figure 2 Mathematical formula fitting the pangenome and core genome size when the genome number of *Clostridium beijerinckii* strains varied from 1 to 233. The cumulative curve (in blue) indicates an open pangenome.

ammonium (Tao et al. 2014). Genome annotation of the five *C. beijerinckii* strains from PM revealed that complete metabolic and transport pathways of many sugars were predicted to be present in all five strains. Furthermore, all five strains were predicted to be able to convert starch, dextrin, trisaccharides (raffinose and manninotriose), disaccharides (sucrose, melibiose, epimelibiose, maltose, trehalose, cellobiose, and isomaltose), monosaccharides (fructose, mannose, galactose, xylose, xylulose, and ribose), organic acids (L-gulonate, D-glucuronate, D-tagaturonate, D-fructuronate, D-galacturonate, D-glycerate, and 4-aminobutanoate), alcohol (mannitol and galactitol), arbutin, salicin, and N-acetyl-D-glucosamine into glucose or intermediates of glycolysis, which subsequently enter into the central carbon metabolism pathway to produce biomass or energy. In addition, the predicted transport systems for these substrates were annotated by KEGG, including ATP-binding cassette (ABC) transporters and phosphotransferase (PTS) systems. Sugars, including glucose, fructose, maltose, sucrose, cellobiose, trehalose, mannose, mannitol, galactitol, and N-acetyl-D-glucosamine, are mainly transported into the cell via the PTS system, but others are transported in part by ABC transporters, such as raffinose, isomaltotriose, maltotriose, melibiose, ribose, D-xylose, arbutin, and salicin. In addition, D-xylose and melibiose can be transported by the D-xylose proton-symporter XylT and the Na⁺/melibiose symporter, respectively. The presence of transporters and the complete metabolic pathways of various substrates indicates that the five *C. beijerinckii* strains likely assimilate various substrates in PM, allowing them to adapt to the environment of PM of SFB ecosystems.

We reconstructed the biosynthetic pathways of ABE and IBE from glucose (Figure 1). All five *C. beijerinckii* strains appear to possess the complete biosynthetic pathways of formate, acetate, and ethanol. However, although formate pathways were not detected in strains 3-8 or G3-3 in our previous study (Tian et al. 2019), we identified five and six gene copies encoding pyruvate formate-lyase (EC: 2.3.1.54), which catalyzes the formation of formate from pyruvate, in the genomes of strains 3-8 and G3-3, respectively. In addition, formate efflux transporters were also identified in the five strains. Formate may be transformed into

formyltetrahydrofolate by tetrahydrofolate catalyzed by formate-tetrahydrofolate ligase (EC: 6.3.4.3), which was found in all five strains. *C. beijerinckii* strains 2-1, G3-1, G3-3, and G3-5 possessed the complete biosynthetic pathways of butyrate and butanol, but the gene encoding 3-hydroxybutyryl-CoA dehydrogenase (paaH, EC 1.1.1.157) was not annotated in the genome of strain 3-8. However, strain 3-8 was shown to be capable of producing butyrate with a yield of 6.81g L⁻¹ in our previous report (Tian et al. 2019). This inconsistency between experimental results and genome annotation calls for further examination of the annotation of this gene or experimental validation.

Unlike other *C. beijerinckii* strains used in industrial biotechnology, the five *C. beijerinckii* strains isolated from PM are not currently predicted to be able to produce acetone or isopropanol due to the absence of acetoacetate decarboxylase (Adc, EC : 4.1.1.4), which converts acetoacetate into acetone. This feature may be beneficial for the further application of these strains in SFB manufacturing, as isopropanol is an undesirable flavor compound that causes dizziness and sleepiness when its content is above threshold values (Minqian et al. 2020). In addition, acetone is also an undesired compound that causes SFB to have a peppery taste.

The organic acids produced by the five *C. beijerinckii* strains, including ethyl acetate, ethyl butanoate, and ethyl lactate, may act as precursors of the main flavor compounds of SFB. All five *C. beijerinckii* strains produced ethyl butanoate, one of the representative flavor compounds in SFB, with a yield of 38–51 mg L⁻¹ (Zheng and Han 2016). In addition, genes predicted to encode two esterases, carboxylesterase NA and carboxyl esterase, a/b hydrolase, were found in the genomes of all five strains and may contribute to the biosynthesis of ethyl butanoate.

Clostridium beijerinckii has been used for the production of hydrogen (Fonseca et al. 2020). In PM of SFB, hydrogen also plays important roles in maintaining the stability of the microbial community of the PM ecosystem (Zou et al. 2018a). The hydrogen biosynthetic pathway in *C. beijerinckii* is predicted to be the same as that in *Clostridium butyricum* and involves ferredoxin hydrogenase (HydA, EC: 1.12.7.2), ferredoxin, and nitrogenase (EC: 1.18.6.1) (Zou et al. 2021). These three genes were all found in the five *C. beijerinckii* strains from SFB ecosystems, including 21 copies of the gene encoding ferredoxin that were present in each genome.

Ammonium is present at high levels in PM of SFB ecosystems, with concentrations ranging from 1.86 to 4.2 g kg⁻¹ in PM of different ages (Tao et al. 2014). Two genes coding ammonium transporters, as well as a gene involved in its assimilation pathway, were found in all five *C. beijerinckii* strains. Ammonium is assimilated mainly by glutamine synthetase (EC: 6.3.1.2), which adds ammonium to glutamate for glutamine biosynthesis. Other enzymes catalyzing the assimilation of ammonia include aspartate-ammonia ligase (EC: 6.3.1.1) and asparagine synthase (EC: 6.3.5.4).

The pangenome and core genome of *Clostridium beijerinckii*

A total of 233 *C. beijerinckii* genomes were used for pangenomic analysis (Supplementary Table S3). The average genome size of the 233 *C. beijerinckii* genomes was 6.11 Mb, and the average number of proteins was 5182. The genes in the 233 *C. beijerinckii* genomes were grouped into 12,572 gene clusters. Among these,

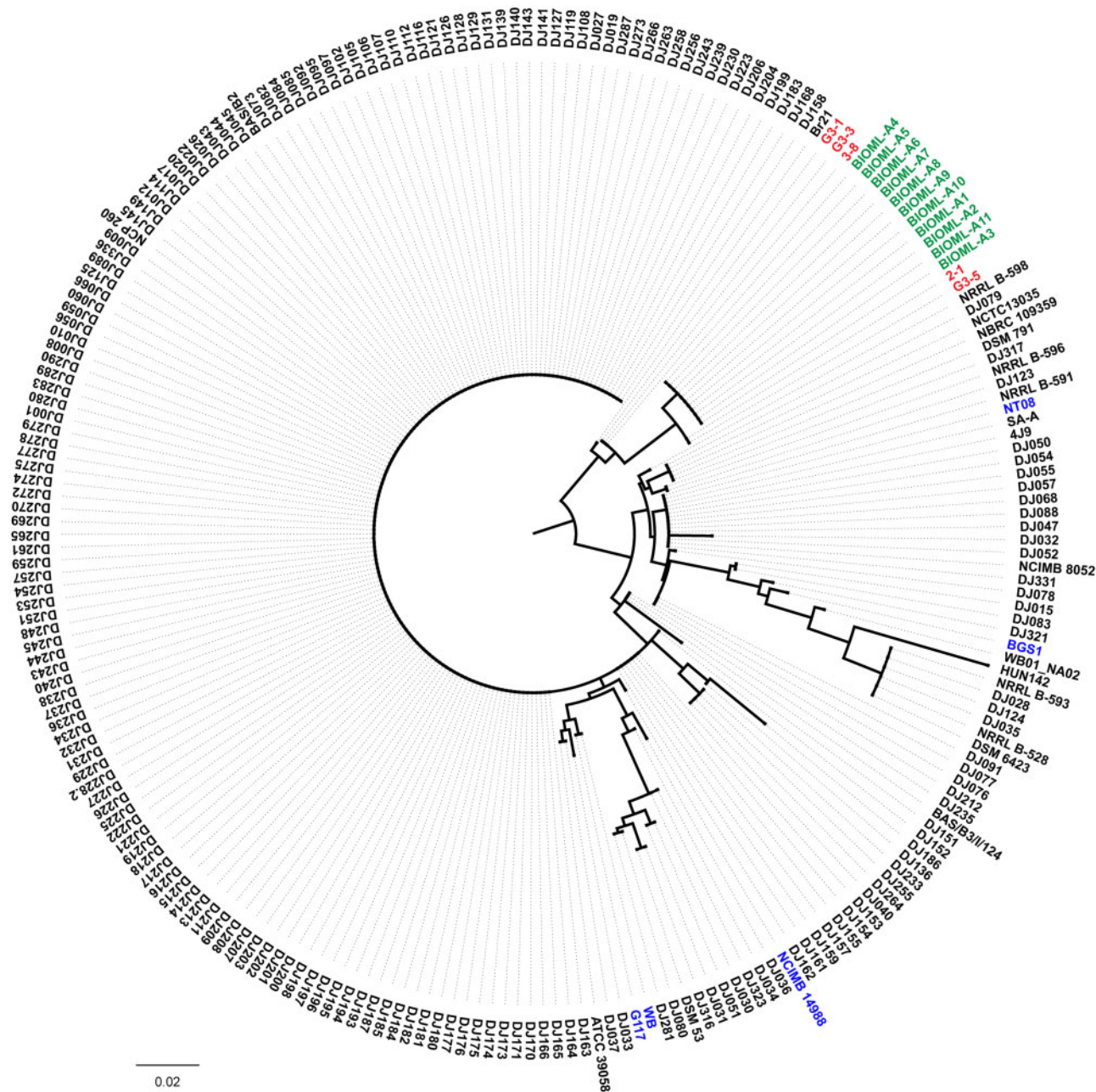


Figure 3 Phylogenetic trees of *Clostridium beijerinckii* strains based on concatenated amino acid sequences of the core genome. Red: strain from pit mud (PM); green: strains from fecal material; blue: strains from soil.

1567 gene clusters were shared among all 233 genomes and made up the core *C. beijerinckii* genome (Supplementary Table S3). The genes in the core genome account for 30.2% of the average number of genes, which is greater than that of *C. butyricum* (26.3%) (Zou et al. 2021). In addition, we found 8851 gene clusters present in two or more, but not all, of the genomes studied that were included in the accessory genome. The average number of accessory genes for each genome was 3153, representing 60.8% of the average number of genes for each genome (Supplementary Table S3). A total of 2154 gene clusters were found in only one genome (strain-specific genes). A total of 179 *C. beijerinckii* strains had no strain-specific gene clusters. Forty-one *C. beijerinckii* strains had between 1 and 10 strain-specific genes in each genome, and 13 strains had more than 10 strain-specific genes.

Strains HUN142 and NCIMB 14988, isolated from rumen contents and garden soil, had the largest number of strain-specific genes, with 387 and 312, respectively (Supplementary Table S3). The five *C. beijerinckii* strains isolated from PM of SFB ecosystems had a total of 33 strain-specific genes.

We fit a curve for the pangenome profile using power-law regression based on Heaps' law (Figure 2). The fitted curve was positive, indicating that the pangenome was predicted to be open ($B = 0.2$) on current sampling of genomes. The *C. beijerinckii* pangenome increased in size when new analyzed genomes were added. This result is similar to those of previous studies on members of the genus *Clostridium* (Udaondo et al. 2017), *C. perfringens* (Kiu et al. 2017; Feng et al. 2020), *C. botulinum* (Bhardwaj and Somvanshi 2017), and *C. butyricum* (Zou et al. 2021). However, it should be

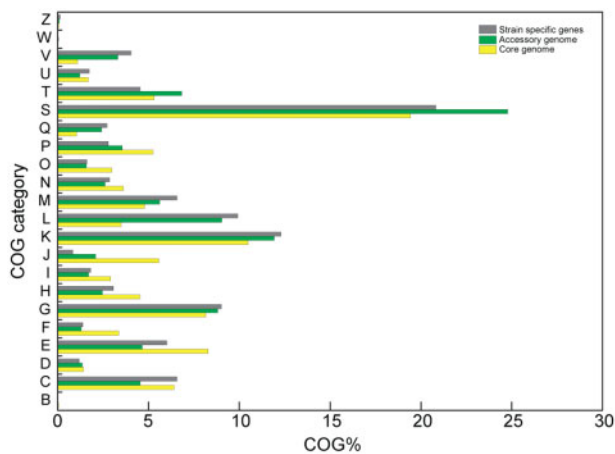


Figure 4 Distribution of Clusters of Orthologous Genes (COG) categories between the core genome, accessory genome, and strain-specific genes of *Clostridium beijerinckii* strains. (B) chromatin structure and dynamics; (C) energy production and conversion; (D) cell cycle control, cell division, chromosome partitioning; (E) amino acid transport and metabolism; (F) nucleotide transport and metabolism; (G) carbohydrate transport and metabolism; (H) coenzyme transport and metabolism; (I) lipid transport and metabolism; (J) translation, ribosomal structure, and biogenesis; (K) transcription; (L) replication, recombination, and repair; (M) cell wall/membrane/envelope biogenesis; (N) cell motility; (O) posttranslational modification, protein turnover, chaperones; (P) inorganic ion transport and metabolism; (Q) secondary metabolite biosynthesis, transport, and catabolism; (S) function unknown; (T) signal transduction mechanisms; (U) intracellular trafficking, secretion, and vesicular transport; (V) defense mechanisms; (W) extracellular structures; and (Z) cytoskeleton.

noted that the pangenome analysis results may be limited by the number of *C. beijerinckii* analyzed genomes. The pangenome profile fitting curve showed that if the analyzed genome number exceeds 4298, the increased pangenome size will be less than one when adding a new genome. Thus, more genomes of *C. beijerinckii* from different ecological niches are expected to be added for further pangenome study.

Phylogenetic analysis of *Clostridium beijerinckii*

The *C. beijerinckii* strains used for pangenomic analysis were isolated from soil, anaerobic sludge, human or pig feces, rumen contents, and PM of SFB ecosystems (Supplementary Table S1). A total of 192 *C. beijerinckii* strains (named with the prefix “DJ”) that have been sequenced are commercial solventogenic Clostridia strains, but detailed information, including their location of isolation and origin, remain unknown. We constructed a phylogenetic tree for the 233 *C. beijerinckii* strains in this study based on concatenated core genome alignments (Figure 3). The five strains isolated from PM of SFB ecosystems were located in the same clade. Strains 3-8, G3-1, and G3-3 were located in a sub-branch close to a separate sub-branch containing strains 2-1 and G3-5, indicating that the five strains were closely related, despite the fact that these five strains were isolated from different pits in different SFB factories. In addition, we found that 11 strains isolated from fecal material were located in a single cluster, which was in the same clade as the five strains from the SFB ecosystems. The five strains isolated from different soil environments were scattered throughout the other clade and showed no phylogeographic relationships.

COG functional annotation of the *Clostridium beijerinckii* pangenome

The genes predicted to be in the pangenome were assigned to COG categories using eggNOG-mapper (Huerta-Cepas et al. 2019). The results showed that 91.1% of the total number of genes in

the core genome, 67.6% of the accessory genome, and 61.7% of the strain-specific genes were assigned COG terms. The largest category was function genes unknown (S), representing 19.4% of the core genome, 24.8% of the accessory genome, and 20.8% of strain-specific genes (Figure 4). Aside from genes assigned to the S category, most genes in the core genome were assigned to transcription (K), amino acid transport and metabolism (E), carbohydrate transport and metabolism (G), or energy production and conversion (C) (Figure 4). COG annotation of accessory genome and strain-specific genes revealed that the largest three categories were similar: K (11.9% and 12.3%, respectively), replication, recombination, and repair (L) (9.0% and 9.9%, respectively), and G (8.8% and 9.0%, respectively). A total of 584 genes in the accessory genome were assigned to the L category, among which were 88 transposases and 49 phage integrase family proteins. However, no transposases or phage-related proteins in the L category were found in the core genome. Similar results were observed in pangenomic analyses of *C. perfringens* (Kiu et al. 2017) and *C. butyricum* (Zou et al. 2021).

Genes shared only by strains isolated from PM of SFB ecosystems

According to the gene pan-matrix obtained from the BPGA pipeline, 298 gene clusters were found only in the five strains isolated from PM, of which 265 belonged to the accessory genome (228 genes shared by all the five strains) and 33 were strain-specific genes. For the 265 genes in the accessory genome, only 73 genes were assigned to COG categories, representing 27.5% of the total number of genes (Table 2). This proportion was much lower than that of the accessory genome (67.6%). COG category distributions showed that the three largest groups were L, S, and K, representing 26.3%, 22.5%, and 11.3% of all genes, respectively. For the L category, five transposase-related genes and four genes belonging to the phage integrase family were found. For metabolic function, three genes, encoding N-acetylmuramoyl-L-alanine amidase, PFAM polysaccharide deacetylase, and Zn peptidase, may play roles in the degradation of macromolecules. N-acetylmuramoyl-L-alanine amidase (EC: 3.5.1.28) hydrolyses the link between N-acetylmuramoyl residues and L-amino acid residues in certain cellwall glycopeptides. PFAM polysaccharide deacetylase has hydrolase activity and acts on carbon-nitrogen (but not peptide) bonds.

For the 33 strain-specific genes, only four genes were assigned to COG categories, all of which were from strain 2-1. The four genes encoded TIGRFAM phage replisome, permease for cytosine/purines, uracil, thiamine, allantoin, PFAM alpha amylase, catalytic, and tRNA cytidyltransferase.

The other 221 genes clusters that were only found in the strains isolated from PM were not assigned to COG categories. This ratio (74.1%) is far lower than those of the accessory genome (32.4%) or strain-specific genes (38.3%). The PM of SFB ecosystems is a relatively closed environment, and the microbial community in PM may have evolved for more than 100 years (based on the age of the PM) (Zhang et al. 2020). Based on the transposase and phage integrase found in the genome, this particular niche may have led to the evolution of the genomes of the five *C. beijerinckii* strains via gene loss or gain from other microbes in the PM of SFB ecosystems.

MGE analysis of five *Clostridium beijerinckii* strains from SFB ecosystems

The average genome size of the 233 sequenced strains was 6.11Mb, which was larger than that of the five stains isolated from PM (5.59Mb), indicating that gene loss events had occurred. In addition, 265 gene clusters were shared only by the strains

Table 2 Clusters of Orthologous Genes (COG) annotation of accessory genes shared only by *Clostridium beijerinckii* strains isolated from pit mud (PM)

COG category	Function description	2-1	3-8	G3-1	G3-3	G3-5
D	Phage tail tape measure protein	1 [#]	1	1	1	1
S	von Willebrand factor, type A	1	1	1	1	1
L	Subunit R is required for both nuclease and ATPase activities, but not for modification	1	1	1	1	1
T	Nacht domain	1	1	1	1	1
K	Bacterial RNA polymerase, alpha chain C terminal domain	1	1	1	1	1
L	DNA primase	1	1	1	1	1
D	DNA recombination	1	1	1	1	1
U	Dynamin family	1	1	1	1	1
S	Dynamin family	1	1	1	1	1
S	Dynamin family	1	1	1	1	1
L	Helicase activity	1	1	1	1	1
M	PFAM Glycosyl transferase family 2	1	1	1	1	1
S	Phage minor structural protein	1	1	1	1	1
L	Domain of unknown function (DUF4277)	1	0 [#]	1	1	1
L	Uncharacterized conserved protein (DUF2075)	1	1	1	1	1
L	TIGRFAM type I restriction system adenine methylase (hsdM)	1	1	1	1	1
EGP	Major facilitator superfamily	1	1	1	1	1
H	Catalyzes the cyclization of GTP to (8S)-3',8-cyclo-7,8-dihydroguanosine 5'-triphosphate	1	1	1	1	1
G	N-Acetylmuramoyl-L-alanine amidase	1	1	1	1	1
K	DNA binding	1	1	1	1	1
V	Type I restriction modification DNA specificity domain	1	1	1	1	1
V	Type I restriction modification DNA specificity domain	1	1	1	1	1
EGP	Major facilitator superfamily	1	1	1	1	1
GM	Methyltransferase FkbM domain	1	1	1	1	1
GM	Methyltransferase FkbM domain	1	1	1	1	1
M	transferase activity, transferring glycosyl groups	1	1	1	1	1
S	Protein of unknown function DUF262	1	1	1	1	1
M	transferase activity, transferring glycosyl groups	1	1	1	1	1
L	Belongs to the "phage" integrase family	1	1	1	1	1
S	PFAM transposase YhgA family protein	1	1	1	1	1
L	Psort location Cytoplasmic, score	1	1	1	1	1
L	Transposase	0	0	1	1	1
K	Bacterial regulatory proteins, tetR family	1	1	1	1	1
K	LysR family	1	1	1	1	1
M	Catalyzes the reduction of dTDP-6-deoxy-L-lyxo-4-hexulose to yield dTDP-L-rhamnose	1	1	1	1	1
S	Glycosyltransferase like family 2	1	1	1	1	1
D	Cell division	1	1	1	1	1
S	Protein of unknown function (DUF2971)	1	1	1	1	1
S	PD-(D/E)XK nuclease family transposase	1	1	1	1	1
S	PFAM Abortive infection protein	1	1	1	1	1
K	PFAM Helix-turn-helix	1	1	1	1	1
E	Pfam: DUF955	1	1	1	1	1
S	head morphogenesis protein, SPP1 gp7 family	1	1	1	1	1
KT	Lecithin retinol acyltransferase	1	1	1	1	1
T	Diguanylate cyclase	1	1	1	1	1
L	Transposase DDE domain	1	1	1	1	1
M	Cell wall binding	1	1	1	1	1
D	Cell wall binding repeat	1	1	1	1	1
L	Transposase DDE domain	1	1	1	1	1
L	Resolvase, N terminal domain	1	1	1	1	1
S	Putative restriction endonuclease	1	1	1	1	1
E	Zn peptidase	1	1	1	1	1
S	Protein of unknown function (DUF2691)	1	1	1	1	1
G	PFAM Polysaccharide deacetylase	0	1	1	1	1

(continued)

Table 2. (continued)

COG category	Function description	2-1	3-8	G3-1	G3-3	G3-5
S	NADPH-dependent FMN reductase	1	1	1	1	1
S	Helix-turn-helix domain	1	1	1	1	1
S	Protein of unknown function (DUF3268)	1	1	1	1	1
S	Domain of unknown function (DUF4258)	1	1	1	1	1
L	Belongs to the "phage" integrase family	1	1	1	1	1
L	Psort location Cytoplasmic, score 8.87	1	1	1	1	1
L	Psort location Cytoplasmic, score	1	1	1	1	1
L	Staphylococcal protein of unknown function (DUF960)	1	1	1	1	1
L	Transposase	1	0	1	1	1
L	Belongs to the "phage" integrase family	1	1	1	1	1
L	Belongs to the "phage" integrase family	1	1	1	1	1
L	Psort location Cytoplasmic, score	1	1	1	1	1
K	Helix-turn-helix XRE-family like proteins	1	1	1	1	1
K	Helix-turn-helix XRE-family like proteins	1	1	1	1	1
K	PFAM helix-turn-helix HxlR type	1	1	1	1	1
S	Domain of unknown function (DUF3797)	1	1	1	1	1
C	Electron transfer flavoprotein	0	0	1	0	1
V	Mate efflux family protein	1	1	1	1	1
L	PFAM transposase, mutator	1	1	1	1	1

#1: exist, 0: not exist.

Table 3 Distribution of genome islands in *Clostridium beijerinckii* strains isolated from pit mud (PM)

Strains	Total size	Number	Total length of GIs/genome size (%)	Total proteins	Hypothetical protein	Phage-related proteins
2-1	406,702	20	7.2	290	155	23
3-8	199,431	19	3.7	232	112	25
G3-1	269,305	21	4.8	290	125	20
G3-3	266,058	21	4.7	290	127	20
G3-5	185,086	18	3.3	230	124	16

from SFB ecosystems, indicating the acquisition of new gene clusters from other microbes in the SFB ecosystem. To investigate MGEs in the five *C. beijerinckii* strains isolated from PM, a total of 15 transposase and 12 phage integrase families were first identified in all five *C. beijerinckii* strains isolated from PM according to COG analysis. Then, GIs were identified in the five *C. beijerinckii* genomes. The number of predicted GIs in the five *C. beijerinckii* genomes ranged from 18 to 21, and the ratios of the total size of GIs to the corresponding genome were between 3.3% and 7.2% (Table 3). Strain 2-1 had the largest total size of GIs, 406,702 bp, representing 7.2% of its total genome. More than 43% of genes identified in GIs were hypothetical proteins with unknown functions. In addition, bacteriophages were identified in the five strains from SFB ecosystems using PHASTER (Arndt et al. 2016). A total of 33 bacteriophage sequences were found, of which 10 were questionable and 23 were incomplete (Supplementary Table S4). Twenty-four bacteriophage sequences were obtained from strains in the genus *Clostridium*, including *Clostridium* phage phiCT19406A (5), *Clostridium* phage phiCT19406B (2), *Clostridium* phage phiCT453A (5), *Clostridium* phage phiCT453B (6), *Clostridium* phage phiCT9441A (4), and *Clostridium* phage phiCTC2B (2). The other nine bacteriophage sequences belonged to *Bacillus* phage BM5 (5) and *Thermoanaerobacterium* phage THSA-485A (4). The existence of a large number of MGEs, including transposases, phage integrases, GIs, and bacteriophage sequences, indicated that the genomes of the five *C. beijerinckii* strains isolated from PM evolved

via the loss or insertion of DNA fragments from the microbiota of SFB ecosystems.

Conclusions

In this study, we analyzed the genomes of five *C. beijerinckii* strains isolated from PM. Metabolic capabilities that are beneficial in the PM environment include assimilation of various carbon sources and ammonium; production of acetate, butyrate, lactate, hydrogen, and esters; and an inability to produce the undesirable flavors isopropanol and acetone. Our analysis of the genomes of 233 *C. beijerinckii* strains revealed an open pangenome with 12,572 gene clusters. A total of 298 genes were found only in the five *C. beijerinckii* strains isolated from PM, among which only 77 genes were assigned to COG categories. The existence of many MGEs indicated that the genomes of the five *C. beijerinckii* strains from the SFB ecosystem evolved in PM. This study will be helpful for future genetic diversification studies and further exploration of *C. beijerinckii* strains isolated from PM.

Data availability

In the present study, strains are available upon request. The raw sequencing reads of *C. beijerinckii* strains 3-8, G3-1, G3-3, and G3-5 were deposited in to Sequence Read Archive (SRA) under accession numbers SRR15304600, SRR15316868, SRR15316869, and SRR15318628.

The genomes of *C. beijerinckii* strains 3-8, G3-1, G3-3, and G3-5 were deposited into GenBank under accession numbers PRJNA690962, PRJNA695099, PRJNA695100, and PRJNA698586, respectively. The genome sequence of *Clostridium beijerinckii* strain 2-1 was deposited into NCBI under accession number PRJNA428897. The other *C. beijerinckii* genome sequences utilized for the pangenome analysis are listed in [Supplementary Table S1](#). All sequences were downloaded from the National Center for Biotechnology Information (NCBI) genome database.

[Supplementary material](#) is available at G3 online.

Acknowledgments

The authors thank International Science Editing (<http://www.internationalscienceediting.com>) for editing this manuscript.

Funding

This work was supported by grants from the National Natural Science Foundation of China (31801522), Sichuan Academician (Expert) Workstation of Solid State Brewing (2017YSGZZ03), and the Open Project Program of Beijing Key Laboratory of Flavor Chemistry, Beijing Technology and Business University (BTBU), Beijing, China (SPFW2020YB08).

Conflict of interest

The authors declare that there is no conflict of interest.

Literature cited

- Arndt D, Grant JR, Marcu A, Sajed T, Pon A, et al. 2016. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* 44:W16–W21.
- Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, et al. 2008. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics.* 9:75.
- Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 19:455–477.
- Bellido C, Loureiro Pinto M, Coca M, González-Benito G, García-Cubero MT. 2014. Acetone-butanol-ethanol (ABE) production by *Clostridium beijerinckii* from wheat straw hydrolysates: efficient use of penta and hexa carbohydrates. *Bioresour Technol.* 167:198–205.
- Bertelli C, Laird MR, Williams KP, Lau BY, Hoad G, et al.; Simon Fraser University Research Computing Group. 2017. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res.* 45:W30–W35.
- Bhardwaj T, Somvanshi P. 2017. Pan-genome analysis of *Clostridium botulinum* reveals unique targets for drug development. *Gene.* 623:48–62.
- Chai LJ, Xu PX, Qian W, Zhang XJ, Ma J, et al. 2019. Profiling the *Clostridia* with butyrate-producing potential in the mud of Chinese liquor fermentation cellar. *Int J Food Microbiol.* 297:41–50.
- Chaudhari NM, Gupta VK, Dutta C. 2016. BPGA- an ultra-fast pan-genome analysis pipeline. *Sci Rep.* 6:24373.
- Coil D, Jospin G, Darling AE. 2015. A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. *Bioinformatics.* 31:587–589.
- Dalal J, Das M, Joy S, Yama M, Rawat J. 2019. Efficient isopropanol-butanol (IB) fermentation of rice straw hydrolysate by a newly isolated *Clostridium beijerinckii* strain C-01. *Biomass Bioenergy.* 127:105292.
- dos Santos Vieira CF, Mauger Filho F, Maciel Filho R, Mariano AP. 2020. Isopropanol-butanol-ethanol (IBE) production in repeated-batch cultivation of *Clostridium beijerinckii* DSM 6423 immobilized on sugarcane bagasse. *Fuel.* 263:116708.
- Drahokoupil M, Patakova P. 2020. Production of butyric acid at constant pH by a solventogenic strain of *Clostridium beijerinckii*. *Czech J Food Sci.* 38:185–191.
- Du Y, Zou W, Zhang K, Ye G, Yang J. 2020. Advances and applications of *Clostridium* co-culture systems in biotechnology. *Front Microbiol.* 11:560223.
- Fang D, Wen Z, Lu M, Li A, Ma Y, et al. 2020. Metabolic and process engineering of *Clostridium beijerinckii* for butyl acetate production in one step. *J Agric Food Chem.* 68:9475–9487.
- Feng Y, Fan X, Zhu L, Yang X, Liu Y, et al. 2020. Phylogenetic and genomic analysis reveals high genomic openness and genetic diversity of *Clostridium perfringens*. *Microb Genom.* 6:mgen000441.
- Fonseca BC, Dalbelo G, Gelli VC, Carli S, Meleiro LP, et al. 2020. Use of algae biomass obtained by single-step mild acid hydrolysis in hydrogen production by the beta-glucosidase-producing *Clostridium beijerinckii* Br21. *Waste Biomass Valor.* 11:1393–1402.
- Golicz AA, Bayer PE, Bhalla PL, Batley J, Edwards D. 2020. Pangenomics comes of age: from bacteria to plant and animal applications. *Trends Genet.* 36:132–145.
- Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 59:307–321.
- Hassan EA, Abd-Alla MH, Zohri A-NA, Ragaey MM, Ali SM. 2019. Production of butanol and polyhydroxyalkanoate from industrial waste by *Clostridium beijerinckii* ASU10. *Int J Energy Res.* 43:3640–3652.
- Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, et al. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47:D309–D314.
- Jiang J, Yang B, Ross RP, Stanton C, Zhao J, et al. 2020. Comparative genomics of *Pediococcus pentosaceus* isolated from different niches reveals genetic diversity in carbohydrate metabolism and immune system. *Front Microbiol.* 11:253.
- Jin G, Zhu Y, Xu Y. 2017. Mystery behind Chinese liquor fermentation. *Trends Food Sci Technol.* 63:18–28.
- Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform.* 20:1160–1166.
- Kiu R, Caim S, Alexander S, Pachori P, Hall LJ. 2017. Probing genomic aspects of the multi-host pathogen *Clostridium perfringens* reveals significant pangenome diversity, and a diverse array of virulence factors. *Front Microbiol.* 8:2485.
- Li HG, Luo W, Gu QY, Wang Q, Hu WJ, et al. 2013. Acetone, butanol, and ethanol production from cane molasses using *Clostridium beijerinckii* mutant obtained by combined low-energy ion beam implantation and N-methyl-N-nitro-N-nitrosoguanidine induction. *Bioresour Technol.* 137:254–260.
- Liu HL, Sun BG. 2018. Effect of fermentation processing on the flavor of Baijiu. *J Agric Food Chem.* 66:5425–5432.
- Luo X, Li H, Sun X, Zheng F, Chai LJ, et al. 2019. Zooming in on butyrate-producing *Clostridial* consortia in the fermented grains

- of Baijiu via gene sequence-guided microbial isolation. *Front Microbiol.* 10:1397.
- Luo R, Liu B, Xie Y, Li Z, Huang W, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience.* 1:18.
- Minqian Z, Gongliang L, Yongtao F, Weidong B, Lianzhong A, et al. 2020. Research progress on control technology of fusel oil during Baijiu fermentation. *China Brew.* 39:8–12.
- Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 35:W182–W185.
- Munch-Petersen E, Boundy CA. 1963. Bacterial content in samples from different sites in the rumen of sheep and cows as determined in two culture media. *Appl Microbiol.* 11:190–195.
- Qureshi N, Blaschek HP. 2001. Recent advances in ABE fermentation: hyper-butanol producing *Clostridium beijerinckii* BA101. *J Ind Microbiol Biotechnol.* 27:287–291.
- Reddy LV, Veda AS, Wee YJ. 2020. Utilization of banana crop residue as an agricultural bioresource for the production of acetone-butanol-ethanol by *Clostridium beijerinckii* YVU1. *Lett Appl Microbiol.* 70:36–41.
- Schubert M, Lindgreen S, Orlando L. 2016. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res Notes.* 9:88.
- Seelert T, Ghosh D, Yargeau V. 2015. Improving biohydrogen production using *Clostridium beijerinckii* immobilized with magnetite nanoparticles. *Appl Microbiol Biotechnol.* 99:4107–4116.
- Survase SA, Jurgens G, van Heiningen A, Granstrom T. 2011. Continuous production of isopropanol and butanol using *Clostridium beijerinckii* DSM 6423. *Appl Microbiol Biotechnol.* 91:1305–1313.
- Tao Y, Li J, Rui J, Xu Z, Zhou Y, et al. 2014. Prokaryotic communities in pit mud from different-aged cellars used for the production of Chinese strong-flavored liquor. *Appl Environ Microbiol.* 80:2254–2260.
- Tian Y, Heng X, Zou W, Ye G. 2019. Isolation and identification of clostridia from the pit mud of Strong-flavor Baijiu and comparative study on butyric acid production. *Food Ferment Ind.* 45:60–65.
- Udaondo Z, Duque E, Ramos JL. 2017. The pangenome of the genus *Clostridium*. *Environ Microbiol.* 19:2588–2603.
- Vernikos G, Medini D, Riley DR, Tettelin H. 2015. Ten years of pan-genome analyses. *Curr Opin Microbiol.* 23:148–154.
- Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One.* 9:e112963.
- Wischril D, Zhang J, Cheng C, Lin M, De Souza LMG, et al. 2016. Production of 1,3-propanediol by *Clostridium beijerinckii* DSM 791 from crude glycerol and corn steep liquor: process optimization and metabolic engineering. *Bioresour Technol.* 212:100–110.
- Xu Y, Wang D, Fan WL, Mu XQ, Chen J. 2010. Traditional Chinese biotechnology. *Adv Biochem Eng Biotechnol.* 122:189–233.
- Xu Y, Zhu Y, Li X, Sun B. 2020. Dynamic balancing of intestinal short-chain fatty acids: the crucial role of bacterial metabolism. *Trends Food Sci Technol.* 100:118–130.
- Yao F, Yi B, Shen C, Tao F, Liu Y, et al. 2015. Chemical analysis of the Chinese Liquor Luzhoulaojiao by comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry. *Sci Rep.* 5:9553.
- Zhang H, Meng Y, Wang Y, Zhou Q, Li A, et al. 2020. Prokaryotic communities in multidimensional bottom-pit-mud from old and young pits used for the production of Chinese strong-flavor baijiu. *Food Chem.* 312:126084.
- Zhao Y, Jia X, Yang J, Ling Y, Zhang Z, et al. 2014. PanGP: a tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics.* 30:1297–1299.
- Zheng X-W, Han B-Z. 2016. Baijiu (白酒), Chinese liquor: history, classification and manufacture. *J Ethn Foods.* 3:19–25.
- Zhong Z, Zhang W, Song Y, Liu W, Xu H, et al. 2017. Comparative genomic analysis of the genus *Enterococcus*. *Microbiol Res.* 196:95–105.
- Zou W, Ye G, Zhang K. 2018a. Diversity, function, and application of *Clostridium* in Chinese strong flavor baijiu ecosystem: a review. *J Food Sci.* 83:1193–1199.
- Zou W, Ye G, Zhang K, Yang H, Yang J. 2021. Analysis of the core genome and pangenome of *Clostridium butyricum*. *Genome.* 64:51–61.
- Zou W, Zhao C, Luo H. 2018b. Diversity and function of microbial community in Chinese strong-flavor Baijiu ecosystem: a review. *Front Microbiol.* 9:671.

Communicating editor: D. Baltrus