Food Research International 151 (2022) 110904

Contents lists available at ScienceDirect

Food Research International

journal homepage: www.elsevier.com/locate/foodres

Shotgun metagenomics of *Cheonggukjang*, a fermented soybean food of Korea: Community structure, predictive functionalities and amino acids profile

Jyoti Prakash Tamang[a,*], Souvik Das[a], Pynhunlang Kharnaior[a], Priyambada Pariyar[a], Namrata Thapa[b,**], Seung-Wha Jo[c], Eun-Jung Yim[c], Dong-Hwa Shin[d]

[a] *DAICENTER (DBT-AIST International Centre for Translational and Environmental Research) and Bioinformatics Centre, Department of Microbiology, School of Life Sciences, Sikkim University, Gangtok 737102, Sikkim, India*
[b] *Biotech Hub, Department of Zoology, Nar Bahadur Bhandari Degree College, Sikkim University, Tadong 737102, Sikkim, India*
[c] *Microbial Institute for Fermentation Industry (MIFI), Sunchang 56048, Republic of Korea*
[d] *Shindonghwa Food Research Institute, Seoul 06192, Republic of Korea*

ARTICLE INFO

Keywords:
Cheonggukjang
Microbial community
Shotgun sequencing
Predictive functionality
Amino acid profiles

ABSTRACT

Cheonggukjang is a naturally fermented soybean food of Korea. The present study was aimed to reveal the whole microbial community structure of naturally fermented *cheonggukjang* along with the prediction of microbial functional profiles by shotgun metagenomic sequence analysis. Metataxonomic profile of *cheonggukjang* samples showed different domains viz. bacteria (95.83%), virus (2.26%), unclassified (1.84%), eukaryotes (0.05%) and archaea (0.005%). Overall, 44 phyla, 286 families, 722 genera and 1437 species were identified. *Firmicutes* was the most abundant phylum (98.04%) followed by *Proteobacteria* (1.49%), *Deinococcus-Thermus* (0.14%). *Bacillus thermoamylovorans* was the most abundant species in *cheonggukjang* followed by *Bacillus licheniformis*, *Bacillus glycinifermentans*, *Bacillus subtilis*, *Bacillus paralicheniformis*, *Bacillus amyloliquifaciens*, *Brevibacillus borstelensis*, *Brevibacillus sonorensis Brevibacillus, Acinetobacter, Carnobacterium, Paenibacillus, Cronobacter Enterococcus, Enterobacter, Terriglobus, Psychrobacter* and *Virgibacillus*. A colossal diversity of the genus *Bacillus* was detected with 150 species. Functional analysis of *cheonggukjang* metagenome revealed the genes for the synthesis and metabolism of wide range of bioactive compounds including, various essential amino acids, conjugated amino acids, different vitamins, flavonoids, and enzymes. Amino acid profiles obtained from KEGG annotation in *cheonggukjang* were validated with experimental result of amino acid profiles.

1. Introduction

Consumption of traditional fermented soybean foods has a long dietary history in the Korean gastronomy since 4000 years ago (Lee & Kim, 2016). Korean fermented soybean foods are of two types: mould-fermented soybean foods such as *doenjang* (sauce, used as condiment), *gochujang* (paste, used as soup), *kanjang* (red pepper paste, used as pickle) (Shin, Kwon, Kim, & Jeong, 2012), *meju*, non-food soya-based starter like *koji* in Japan for fermentation of mould-oriented fermented soybean foods (Shin, & Jeong, 2015), and bacterial-fermented soybean food: *cheonggukjang,* differently spelled as *chongkukjang/cheonkokjang* (sticky, non-salted fermented whole soybean food, eaten as gravy/side-dish) (Shin, Kim, Park, & Kim, 2016). *Cheonggukjang* is believed to be originated around 1st century BCE (Kwon, Chung, & Jang, 2019), and was considered as one of the most delicious food items during Koryo dynasty in Korea (935–1392 CE) (Kwon, Chung, & Jang, 2019). During traditional method of preparation of *cheonggukjang*, soybeans are boiled, the excess water is discarded, and cooked beans are wrapped in paddy straw, and placed in a warm stone floor near/above earthen kitchen for natural fermentation for 3–4 days (Shin, Kim, Park, & Kim, 2016). The fermented soybeans are covered with whitish viscous sticky materials, presumably poly-glutamic acid (Hsueh, Huang, Kunene, & Lee, 2017), with umami taste (Hartley, Liem, & Keast, 2019) preferred by the Korean consumers, and is eaten as gravy with boiled rice (Kim et al.,

---

* Corresponding author.
** Corresponding author.
E-mail addresses: jptamang@cus.ac.in (J.P. Tamang), dmalati22@hotmail.com (N. Thapa).

https://doi.org/10.1016/j.foodres.2021.110904
Received 21 March 2021; Received in revised form 10 September 2021; Accepted 13 December 2021

2016). Nowadays, *cheonggukjang* is produced in small to large food manufacturing units in Korea by using commercially available starter culture of *Bacillus subtilis* (Lee et al., 2015). However, traditional preparation of naturally fermented *cheonggukjang* is observed only in few villages in Republic of Korea (Kim, Kim, Jeong, Kim, & Shin, 2006). *Cheonggukjang* is similar to other sticky fermented soybean food of Asia such as *kinema* of India, Nepal and Bhutan, *natto* of Japan, *thua nao* of Thailand, *pe poke* of Myanmar and *sieng* of Laos (Tamang, 2015; Tamang, Watanabe, & Holzapfel, 2016a).

Bacillus subtilis is a dominant fermenting bacterium in *cheonggukjang* (Kindoli, Lee, Heo, & Kim, 2012; Moon et al., 2015), followed by other species of *Bacillus* such as *B. velezensis* (Bang et al., 2020), *B. amyloliquefaciens* (Wu & Ahn, 2011), *B. glycinifermentans* (Kim, Dunlap, Kwon, & Rooney, 2015), and lactic acid bacteria such as *Enterococcus faecium, E. durans* and *E. sanguinicola* (Park et al., 2014, 2020). Bacteriophage belonging to *Myoviridae* is also present in *cheonggukjang* (Ghosh, Kang, Hyun, & Kim, 2018). Several functional properties and health benefits of *cheonggukjang* have been reported to consider this food as healthy Korean cultural food (Shin, Kim, Park, & Kim, 2016; Tamang, Shin, Jung, & Chae, 2016b; Kwon, Chung, & Jang, 2019). Some functional properties of *cheonggukjang* are antioxidants (Kim et al., 2011; Ali et al., 2018), probiotics (Park et al., 2014, 2020; Park, Lee, Shim, Yum, & Seo, 2020), immunomodulators (Bae, Shin, See, Chai, & Shon, 2014), anti-haemolytic activity (Song et al., 2010) and bioavailability of secondary metabolites (Baek et al., 2010; Park et al. 2011, 2010; Kim et al. 2012; Kwon, Lee, Lee, Kim, & Lee, 2019; Kim et al., 2020).

The culture-independent methods using the pyrosequencing (Nam, Yi, & Lim, 2012) and DGGE methods (Hong, Choi, & Chung, 2012; Hong, Lim, Kim, Shin, & Chung, 2013) were earlier applied to profile the bacterial community structures in *cheonggukjang*. However, such methods may sequence only short length of nucleotides (Cao, Fanning, Proos, Jordan, & Srikumar, 2017). The profiling of entire microbial community up to species levels can be analysed by the shotgun metagenomic sequencing tool (Xie et al., 2019; Tamang et al., 2020), that may sequence the genomes of untargeted cells in a microbial community to decode community structures including culturable and unculturable bacteria, yeasts, fungi, virus and archaea in food samples (Arıkan, Mitchell, Finn, & Gürel, 2020; Leech et al., 2020). Hence, this paper is aimed to study the metataxonomic of microbial community in naturally fermented *cheonggukjang* of Korea by shotgun metagenomic sequencing method, supported by machine learning tools. Functional profiles of metagenomes were also predicted using the SqueezeMeta pipeline and KEGG database and were validated with amino acid profiles of the samples.

2. Materials and methods

2.1. Collection of samples

A total of 10 samples of naturally fermented *cheonggukjang* were collected from four different provinces of South Korea viz., Gangjim and Damyang regions of Jeonnam province (2 samples), from Chungju region of Chungbuk province (1 sample), Taebaekk and Bonghwa regions of Kangwon province (2 samples) and Jeongeup region of Jeonbuk province (5 samples) by Prof. Shin and his team. Fresh samples, after collection were freeze dried (Weißbecker, Buscot, & Wubet, 2017) for 48 h and grinded into powder form, packed in a pre-sterile plastic tubes, capped, sealed and transported to Prof. Tamang of Department of Microbiology, Sikkim University, India by air-mail parcel. After receiving, the samples were kept in moist-free condition for further analysis, since sun-dried fermented soybean products may be stored at room temperature for several weeks (Tamang, 2015).

2.2. Measurement of viscosity

The dynamic viscosity of *cheonggukjang* was determined using the modified method of Wu et al. (2020). Twenty grams of samples were mixed with 50 mL of distilled water, and were subjected to vigorous shaking in a conical flask (250 mL) for 30 min. The slimy part was collected and 30 mL of its aliquot was measured for dynamic viscosity (100 rpm at 20℃) using a viscometer (DV1MRVTJ0, Brookfield AMETEK, MA, USA) The experiment was done in triplicate sets. The viscosity was also calculated in percentage taking the value of the solvent i.e. distilled water as the highest value for comparison.

2.3. Extraction of metagenomic DNA

Ten grams of samples were homogenised in Stomacher (400 Circulator, Seward, UK) with 90 mL of sterile 0.1 M phosphate buffer saline (pH 6.4) for 5 min. After homogenization, the homogenate was filtered and the filtrate was used for the extraction of genomic DNA using the Nucleospin® Food DNA kit (MACHEREY-NAGEL GmbH & Co. KG, Duran, Germany) as per the manufacturer's protocol. Concentration of DNA was then quantified using spectrophotometer (Eppendorf, USA). The quality of DNA was checked in 0.8% agarose gel electrophoresis and visualized using Gel Doc EZ imager (BioRad, USA). According to the collection sites, the DNA from the samples of four provinces were pooled in equal quantity (Ray et al., 2019), and the mixture was considered as one sample. Ultimately, 4 samples (from 10 initial samples), i.e. CKCB (Chungbuk province), CKJB (Jeonbuk province), CKJN (Jeonnam province) and CKKW (Kangwon province) were used for the further metagenomic and other types of analysis.

2.4. MinION library preparation and nanopore sequencing

The MinION gDNA sequencing Ligation Kit SQK-LSK109 (Oxford Nanopore Technologies, Oxford, United Kingdom) was used to prepare the metagenome libraries according to the manufacturer's instruction. The method of Sevim et al. (2019) was followed for preparation of metagenome libraries of samples. Ten µg of genomic DNA was sheared using the Covaris g-tubes to produce the fragments of the size > 10 kb (Covaris Inc., MA, USA). The sheared genomic DNA with appropriate size was then selected using the BluePippin instrument (Sage Science, Beverly, MA, USA). End-repairing was performed on the sheared DNA using the NEBNext FFPE DNA Repair kit following the manufacturer's protocol (New England BioLabs, Ipswich, MA, USA). Qubit HS DNA kit was used to quantify the gDNA. Ligation Sequencing Kit SQK-LSK109 (Oxford Nanopore Technologies, Oxford, United Kingdom) was used for the clean-up and adapter ligation as per the manufacturer's instruction. The ligated sample was purified and eluted using the AMPure XP beads (Beckman Coulter Inc., USA) along with wash buffer and kit. The libraries were then sequenced (1D sequencing) on MinION MK1 device using R9 flow cell chemistry. The device was controlled through the MinKNOW software version 1.0.5 (Oxford Nanopore Technologies, Oxford, United Kingdom) and Metrichor platform of ONT was used to perform the 1D base calling.

2.5. Bioinformatics analysis

2.5.1. Metataxonomic profiles of metagenome

Raw reads obtained from MinION platform in fast5 format were converted to fastq format using the poretools software version 0.6.0 (Loman & Quinlan, 2014). NanoPlot version 1.30.1 (De Coster, D'Hert, Schultz, Cruts, & Van Broeckhoven, 2018) was used to check the quality of the raw reads and were assembled via the Canu assembler (Koren et al., 2017). Taxonomic assignment of the quality-checked and assembled sequences were performed in Kaiju pipeline version 1.7.4 (Menzel, Ng, & Krogh, 2016) with the default parameters against NCBI BLAST *nr* + *euk* database that contains millions of protein sequences from bacteria, viruses, eukaryotes and archaea (Chen, Huang, & Wu, 2017). Default "greedy" algorithm was preferred to map the sequences against the database (Zhang, Schwartz, Wagner, & Miller, 2000).

Minimum cut-off for match length was maintained 11; however, cut-offs for minimum match score and allowed mismatches were kept 80 and 5, respectively. Within the pipeline, the reads were first translated into the possible open reading frames and the corresponding amino acid sequences were split into fragments based on the stop codons. Then the fragments were sorted according to their BLOSUM62 (BLOcks SUbstitution Matrix) scores (Lazar, Karcini, Ahuja, & Estrada-Palma, 2019). Ultimately the fragments were searched against the database via backwords search approach using the Burrows-Wheeler algorithm (Menzel, Ng, & Krogh, 2016).

2.5.2. Predictive functional profiles

Predictive functional profiles of the metagenome were performed on Quality-filtered contigs using the SqueezeMeta pipeline version 1.3.0 (Tamames & Puente-Sánchez, 2019). Short contigs (<500 bp) were removed using prinseq version 0.20.4 (Schmieder & Edwards, 2011) after importing into the pipeline. Open Reading Frames (ORFs) from the assembled nanopore reads were predicted through Prodigal software version 2.6.3 (Hyatt, LoCascio, Hauser, & Uberbacher, 2012). Diamond software (Buchfink, Xie, & Hutson, 2015) was used for the mapping of gene sequences against different functional databases, such as evolutionary genealogy of genes, non-supervised Orthologous Groups (eggnog) database (Huerta-Cepas et al., 2016) for Clusters of Orthologous Groups/Non-supervised Orthologous Groups (COG/NOG) annotation (Galperin, Kristensen, Makarova, Wolf, & Koonin, 2019) and Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa, & Goto, 2000) for KEGG ID annotation. Pathway prediction was analysed using MinPath (Ye, & Doak, 2009). Best hit approach (Arango-Argoty et al., 2018) with the average bitscore of >20% was followed during the functional assignment. Most of the analysis were performed with the default parameters in different software. The pathways were further subdivided into 3 categories: level 1, level 2 (super-pathway) and level 3 (sub-pathways) (Scala, Serra, Marwah, Saarimäki, & Greco, 2019).

2.6. Statistical analysis

2.6.1. Pooled sequences

Nucleotide diversity (pi) analysis and different indices of neutrality test based on sequence polymorphism in DnaSP software version 6 (Rozas et al., 2017) were performed to justify the pooling of DNA from 10 different samples in terms of intra sample (within the sample) diversity. Among the different indices of neutrality tests, Tajima's D value and Fu's F_s statistics were performed in the same software. To check the intra sample diversity in terms of species level and functional profiles, one sample T test was performed in IBM SPSS version 20 for all the samples. Statistical relations among the samples were performed using Mann-whitney test (Li et al., 2017) in terms of species level and level 3 functional profiles in PAST version 4.0 software (Martino et al., 2019).

2.6.2. Alpha and beta indices

Non-parametric Shannon index and Simpson's index of diversity (1-D) were calculated using PAST software version 4.0 (Martino et al., 2019). Bray-Curtis index of beta diversity was also calculated using PAST version 4.0 and visualised via Principal Coordinates Analysis (PCoA) plot (Martino et al., 2019). Log transformation of the data was performed before constructing the PCoA plot. Fisher's exact test (non-parametric) was performed through STAMP software version 2.1.3 (Parks, Tyson, Hugenholtz, & Beiko, 2014) to check the significance in the distribution of different microbial taxa among the samples.

2.6.3. Functional profiles

Clustering pattern among the samples was checked through PCoA plot using the level 3 sub-pathways. The PCoA plot was constructed after the log transformation of data and then visualised in ClustVis web tool (Metsalu, & Vilo, 2015). Non-parametric Spearman's rank correlation was performed using IBM SPSS software version 20 and PAST version 4.0 (Martino et al., 2019) software to check the correlation between the microorganisms in samples and different levels (super-pathways) functional profiles. Heatmap was constructed via ClustVis web tool to visualize the correlation profiles. Correlation between different bacteria and the amino acid profiles obtained from KEGG annotation was also performed.

2.7. Analysis of free amino acids

About 2 g of each sample was shaken in 50 mL of distilled water for 20 min. The extract was centrifuged at 3000 rpm for 10 min and 2 mL of 5% TCA was added to 2 mL of supernatant. Then the mixture was centrifuged again at 10,000 rpm for 10 min, the supernatant was taken and diluted with 0.02 N-HCL and passed through a 0.2 μm syringe filter. The free amino acid composition in samples was determined with a Hitachi Amino Acid Analyzer (L-8900, Hitachi, Tokyo, Japan) after hydrolysis of 100 mg protein with 6 M HCl at 110 °C for 24 h. An amino acid standard mixture solution for automatic amino acid analysis (Type H, Wako Pure Chemical Industries Ltd., Osaka, Japan) was used for the quantification of endogenous amino acid content. The experiment was performed in triplicate sets.

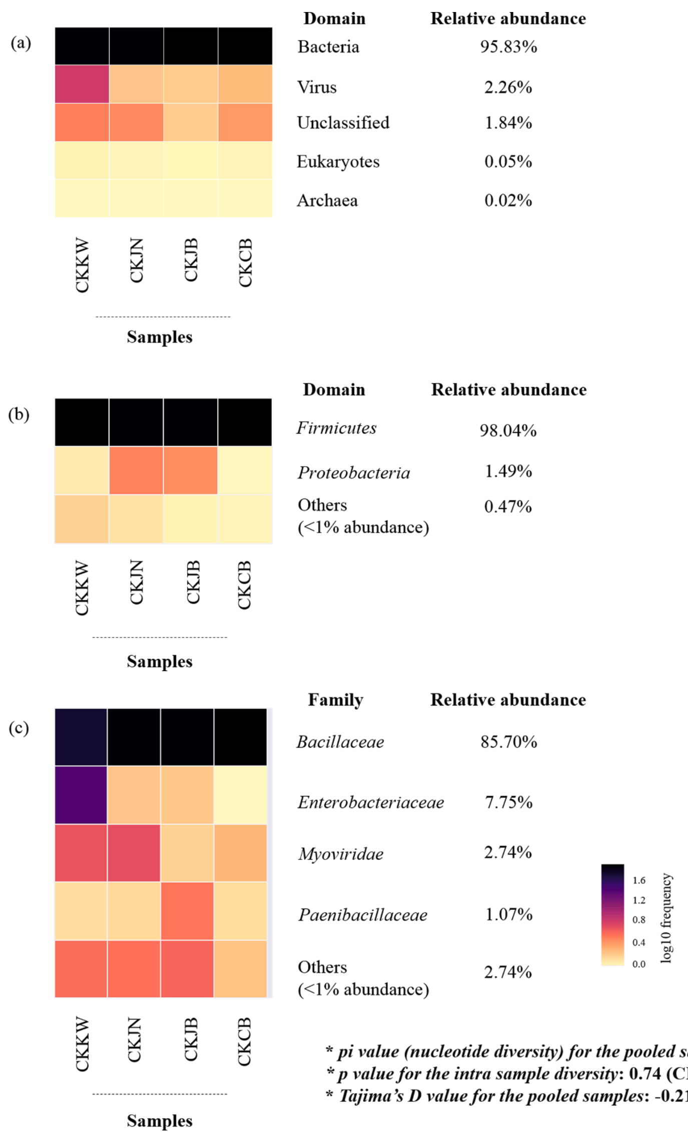
3. Results

3.1. Pool sequence

As mentioned earlier, *cheonggukjang* samples were pooled on the basis of their geographical origin and we hypothesized that the samples from the same origin possess the minimal diversity differences. Nucleotide diversity (pi) per site was 0.712 (CKCB), 0.713 (CKJB), 0.713 (CKJN) and 0.733 (CKKW), respectively. However, the theta values (nucleotide difference) found for each of the samples were 0.75, 0.74, 0.74 and 0.75, respectively. Among the different indices of neutrality test, first we checked the Tajima's D value. Each sample showed the negative value for Tajima's D, which were −0.21 (CKCB), −0.19 (CKJB), −0.19 (CKJN) and −0.09 (CKKW), respectively. The second neutrality test indices was Fu's F_S statistics (F_{ST}); F_{ST} values found for all the samples were −5.30, −3.24, −3.24 and −3.88.

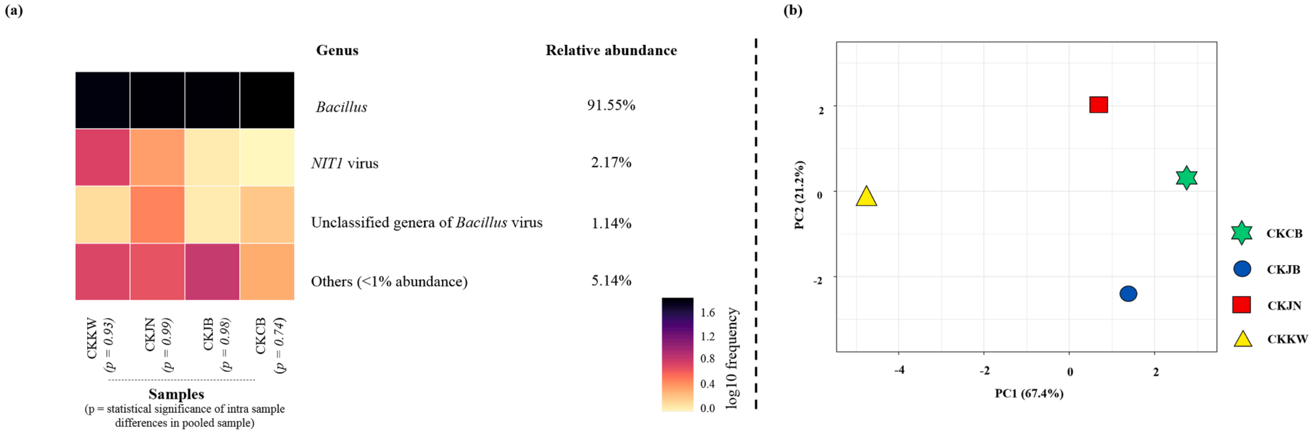
3.2. Metataxonomic profiling

A total of 1,473,066 reads were obtained from samples with an average of 368,266 reads per sample. Average length of the reads was found 1073. Total number of bases recovered from the samples were 278,858,768.0 for CKCB, 608,641,394.0 for CKJB, 367,978,066.0 for CKJN and 254,882,658.0 for CKKW, respectively. Shotgun metagenomic sequence analysis of *cheonggukjang* samples showed different domains viz., bacteria, archaea, viruses and eukaryotes. Bacteria were the abundant domain (95.83%) followed by virus (2.26%), unclassified (1.84%), eukaryotes (0.05%) and archaea (0.02%) (Fig. 1a). Overall, 44 phyla, 286 families, 721 genera and 1439 species were identified from the metagenomic analysis. At phylum level, *Firmicutes* was the most abundant phylum (Fig. 1b) along with the phyla detected at <1% abundance (Supplementary Table 1). No phylum with the abundance of >1% was found from domains of archaea and eukaryotes. Across all the domains, *Bacillaceae* was found the most abundant family followed by *Enterobacteriaceae*, *Myoviridae* [viral family], *Paenibacillaceae* and other minor families (2.74%) (Fig. 1c) with the abundance of <1% (Supplementary Table 2). No family with the abundance of >1% was detected from archaea and eukaryotes. *Bacillus* (91.55%) was the most abundant genus (Fig. 2a) with other minor genera detected at <1% abundance (Supplementary Table 3). No cluster was found among the samples in PCoA analysis according to the abundance/distribution of genera from different domains (Fig. 2b). At species level, *Bacillus thermoamylovorans* was the most abundant species followed by *Bacillus licheniformis*, *Bacillus glycinifermentans*, *Bacillus subtilis*, *Bacillus paralicheniformis*, *Bacillus*



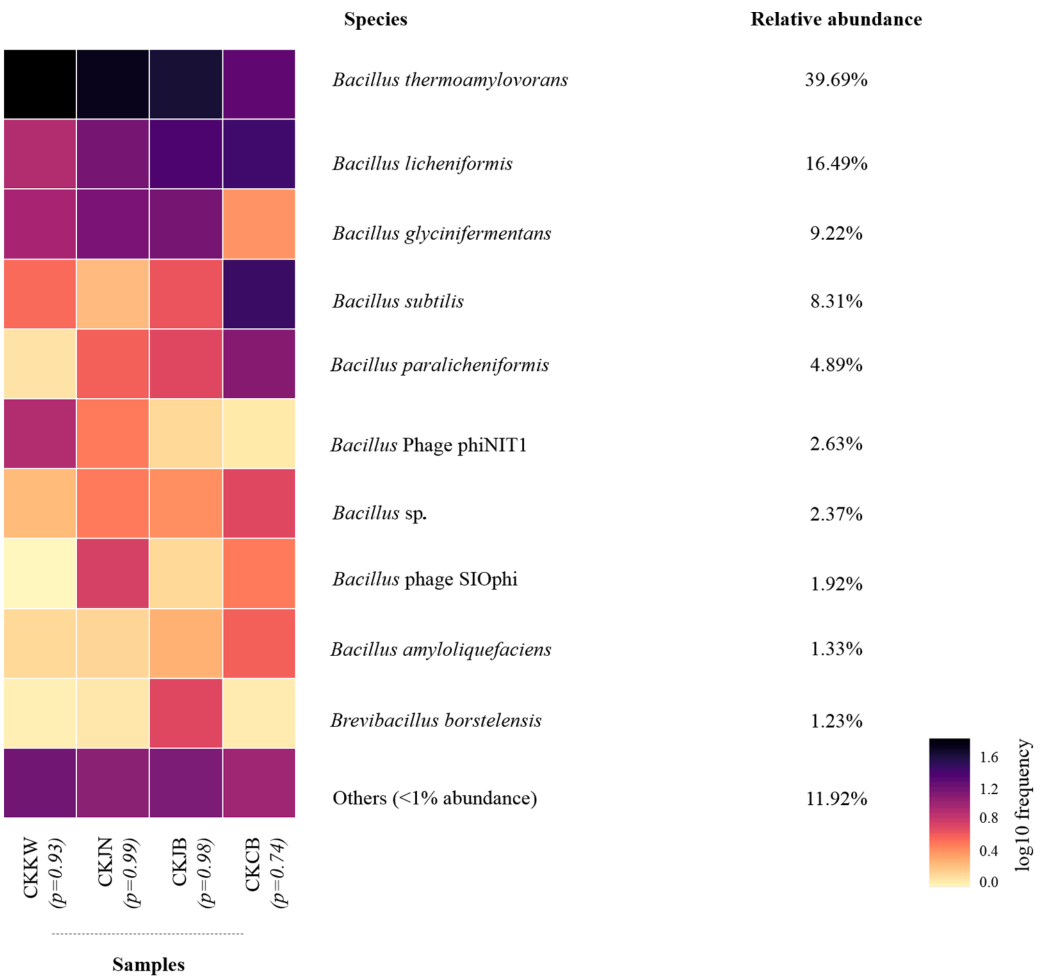
***pi value (nucleotide diversity) for the pooled samples*: 0.712 (CKCB), 0.713 (CKJB), 0.713 (CKJN) and 0.733 (CKKW)**
***p value for the intra sample diversity*: 0.74 (CKCB), 0.98 (CKJB), 0.99 (CKJN) and 0.93 (CKKW)**
***Tajima's D value for the pooled samples*: -0.21 (CKCB), -0.19 (CKJB), -0.19 (CKJN) -0.09 (CKKW)**

Fig. 1. Heatmap showing the relative abundance of different taxa levels identified from the shotgun analysis of *cheonggukjang*: (a) domain (b) phylum and (c) family.



***Pi value (nucleotide diversity) for the pooled samples*: 0.712 (CKCB), 0.713 (CKJB), 0.713 (CKJN) and 0.733 (CKKW)**

Fig. 2. Heatmap and PCoA showing: (a) relative abundance of different genera identified from the shotgun analysis of *cheonggukjang* and (b) relatedness among the samples in terms of genera distribution, respectively.



Species	Relative abundance
Bacillus thermoamylovorans	39.69%
Bacillus licheniformis	16.49%
Bacillus glycinifermentans	9.22%
Bacillus subtilis	8.31%
Bacillus paralicheniformis	4.89%
Bacillus Phage phiNIT1	2.63%
Bacillus sp.	2.37%
Bacillus phage SIOphi	1.92%
Bacillus amyloliquefaciens	1.33%
Brevibacillus borstelensis	1.23%
Others (<1% abundance)	11.92%

*** p values with the sample indicates the intra sample diversity.**
*** *Fu's Fs statistics* ($F_{ST)}$: -5.30 (CKCB), -3.24 (CKJB), -3.24 (CKJN) and -3.88 (CKKW)**
*** *p value for the inter sample diversity among the samples*: 2.148E-41 (CKCB-CKJB), 6.779E-18 (CKCB-CKJN), 1.229E-20 (CKCB-CKKW), 0.008 (CKJB-CKJN), 2.96E-09 (CKJB-CKKW), 0.006 (CKJN-CKKW)**

Fig. 3. Heatmap showing the species level diversity detected in *cheonggukjang*: (a) relative abundance of different species (b) & (c) extended error-bar to show the significant abundance of different species from various taxa level between CKCB-CKKW and CKJB-CKKW, respectively.

amyloliquefaciens, *Brevibacillus borstelensis* and other minor genera (Fig. 3). No species from archaea and eukaryotes were detected at >1% abundance. A total of 150 species of *Bacillus* was detected in *cheonggukjang* metagenome; among them, the genera with >1% abundance were *Bacillus thermoamylovorans*, *Bacillus licheniformis*, *Bacillus glycinifermentans*, *Bacillus subtilis*, *Bacillus paralicheniformis*, *Bacillus* sp. and *Bacillus amyloliquifaciens* (Supplementary Table 4). Lactic acid bacteria (LAB), non-LAB, archaea, yeasts, filamentous moulds, other phages, micro-eukaryotes and parasites were also detected in <1% abundance from *cheonggukjang* samples (Supplementary Table 5).

A total of 53 viral species with 31 *Bacillus* phages and 17 other phages were detected from *cheonggukjang* samples (Supplementary Table 6). From the domain of eukaryotes, 14 algal species, 4 species from Alveolates, 42 species of filamentous moulds, 8 species of yeasts and 12 parasitic species were detected with <1% abundance (Supplementary Table 7). About 10 species of archaea were also detected in samples with <1% abundance (Supplementary Table 8).

In term of species abundance, diversity within the pooled samples and between the samples were calculated using one sample T test and Mann-whitney test (Li et al., 2017), respectively. Significant species diversity was not found within any of the pooled samples. The P values for CKCB, CKJB, CKJN and CKKW were 0.74, 0.98, 0.99 and 0.93, respectively. This clearly proved that there is no significant differences in the sequences of the pooled samples, which can also justify and reconfirm our hypothesis behind the pooling. However, pooled CKCB was found to be significantly different from pooled CKJB (p = 2.148E-41), pooled CKJN (p = 6.779E-18) and pooled (p = 1.229E-20). Pooled CKJB was also found to be differed significantly from pooled CKJN (p 0.008) and pooled CKKW (p = 2.96E-09). Pooled CKJN and pooled CKKW showed significant difference among them (p = 0.006).

3.3. Diversity indices

The average value for the non-parametric Shannon index was 3.44 ± 0.16 (Mean ± SD) among the samples (Table 1). Shannon index value was found highest in CKJB (3.44), followed by CKCB (3.24), CKJN (3.17) and CKKW (3.05). In case of Simpson's index of diversity (1-D), CKCB showed the highest value (0.82), followed by CKJB (0.80), CKJN

Table 1
Different indices of alpha diversity among the four samples of *cheonggukjang*.

Samples	Shannon index	Simpson (1-D)	Goods coverage
CKCB	3.24	0.82	0.99
CKJB	3.44	0.80	0.99
CKJN	3.17	0.73	0.99
CKKW	3.05	0.65	0.99

Table 2
Relative viscosity of *cheonggukjang* samples.

Place of Collection	Sample	Reading in cP (centipoise) 1st	2nd	3rd	Mean of the Dynamic viscosity (cP)	Viscosity percentage (%)
Chungbuk-Chungju	CKCB	4	4	4	4 ± 0	50
Jeonbuk-Jeongeup	CKJB-1	4	8	4	5.33 ± 2.31	66.625
Jeonbuk-Jeongeup	CKJB-2	4	4	4	4 ± 0	50
Jeonbuk-Jeongeup	CKJB-3	4	4	4	4 ± 0	50
Jeonbuk-Jeongeup	CKJB-4	8	8	4	6.66 ± 2.31	83.25
Jeonbuk-Jeongeup	CKJB-5	8	8	8	8 ± 0	100
Jeonnam-Gangjin	CKJN-1	4	4	4	4 ± 0	50
Jeonnam-Damyang	CKJN-2	4	4	4	4 ± 0	50
Kangwon-Taebaekk	CKKW-1	4	4	4	4 ± 0	50
Kangwon-Bonghwa	CKKW-2	4	4	4	4 ± 0	50

(0.73) and CKKW (0.65). Good's coverage was 0.99 ± 0.001 for all the samples (Table 2). Beta diversity among the samples was represented via PCoA plot using the Bray-Curtis index, and showed no significant clustering among the samples (Fig. 4a).

3.4. Shared and unique species

A diverse microbial community of both shared and unique species was found among the four samples of *cheonggukjang* (Fig. 4b). Around 149 core species of bacteria (common to all samples) were detected from the *cheonggukjang* metagenome (Supplementary Table 9a), some of them include *B. clausii*, *Enterococcus faecium*, *B. okuhidensis, B. glycinifermentans*, *B. paralicheniformis*, *B. niacini*, *B. andreraoultii*, *B. pumilis*, *Enterococcus gallinarum*, *Paenibacillus* sp., *B. amyloliquefaciens*, *B. coagulans* and *B. circulans*. A total of 17 and 20 shared bacterial species were found among CKCB-CKJB-CKJN (e.g. *B. subterraneus*, *Brevibacillus reuszeri*, *B. gobiensis*, *B. infantis* and *Virgibacillus chiguensis*) and CKCB-CKJB-CKKW (e.g. *Lactobacillus casei*, *Lysinibacillus massiliensis*, *B. aryabhattai* and *Paenibacillus odorifer*), respectively; 23 bacterial species were shared between CKCB-CKJB; however, 9 common bacterial species were shared between CKCB and CKJN (Supplementary Table 9b) and 89 and 20 bacterial species were shared between CKJB-CKJN and CKJN-CKKW, respectively (Supplementary Table 9c). Unique bacterial species were 77, 321, 161 and 179 in four samples, respectively (Supplementary Table 9d). About 8 core viral species were detected in *cheonggukjang* metagenome which include *Bacillus* phage SPO1, *Bacillus* phage 1102phi1-3, *Bacillus* phage SIOphi, *Bacillus* phage BCP8-2, *Bacillus* phage vB_BanS-Tsamsa, *Bacillus* phage Grass, *Bacillus* phage phiNIT1 and *Bacillus* phage PM1 (Supplementary Table 10a). 8 viral species were shared among CKCB-CKJB-CKJN (e.g. *Bacillus* phage Grass, *Bacillus* phage phiNIT1 and *Bacillus* phage PM1) and 3 viral species were shared among CKCB-CKJB-CKKW (e.g. *Bacillus* phage Shbh1, *Bacillus* phage Mater and *Streptococcus* phage 53) (Supplementary Table 10b). CKCB and CKJN shared 4 species of virus; whereas, 3 and 2 viral species were shared between CKJB-CKJN and CKJN-CKKW, respectively (Supplementary Table 10b). No shared viral species were found between CKCB and CKJB (Supplementary Table 10c). The number of unique viral species in the samples were 2 (CKCB), 13 (CKJB), 6 (CKJN) and 8 (CKKW), respectively (Supplementary Table 10d). *Mucor ambiguous* was the only core species in the domain of eukaryotes in the samples (Supplementary Table 11a). *Toxoplasma gondii* was found to be shared among CKCB-CKJB-CKJN (Supplementary Table 11a). The unique eukaryotic species in CKCB, CKJB, CKJN and CKKW were 9, 16, 16 and 35, respectively (Supplementary Table 11b). No shared species of archaea

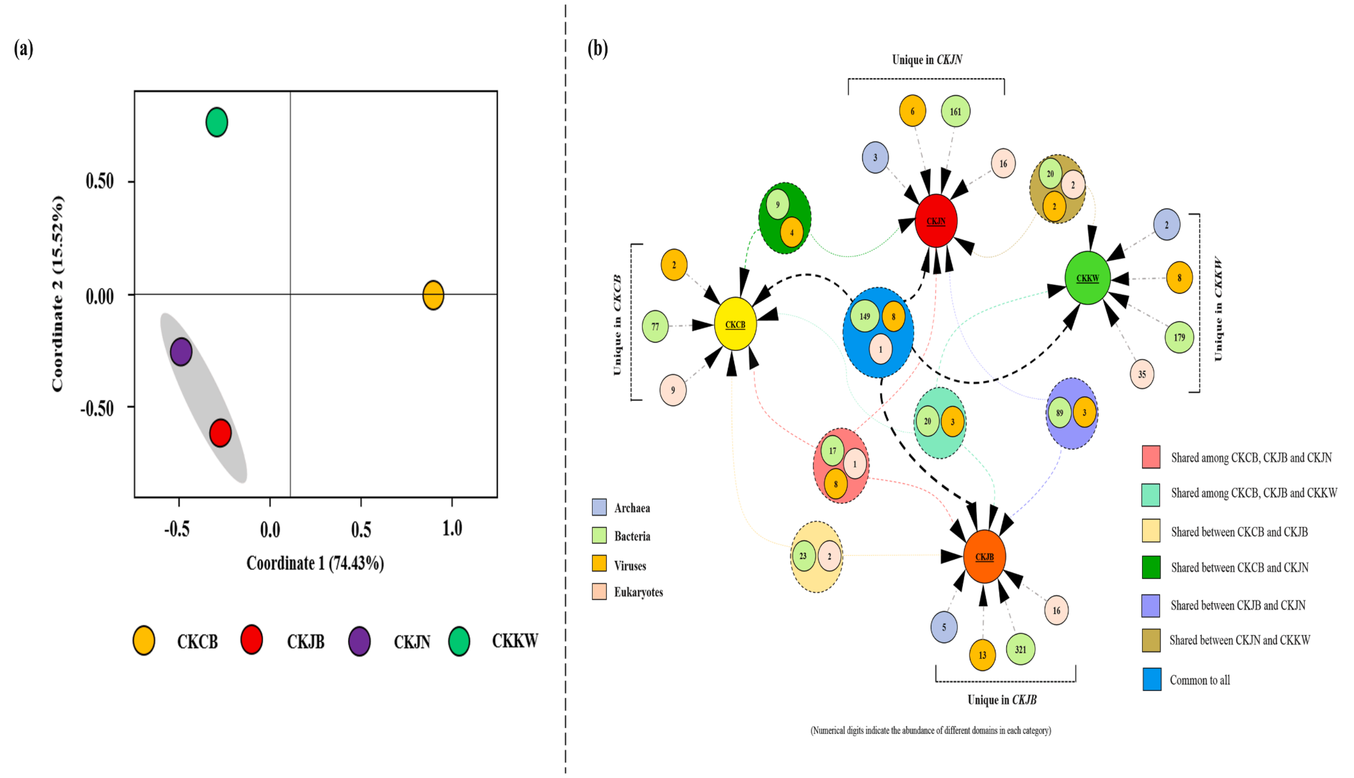


Fig. 4. (a) PCoA plot (based on Bray-Curtis index) showing the difference among the *cheonggukjang* samples in terms of beta diversity, (b) Shared and unique species from different domains among *cheonggukjang* samples.

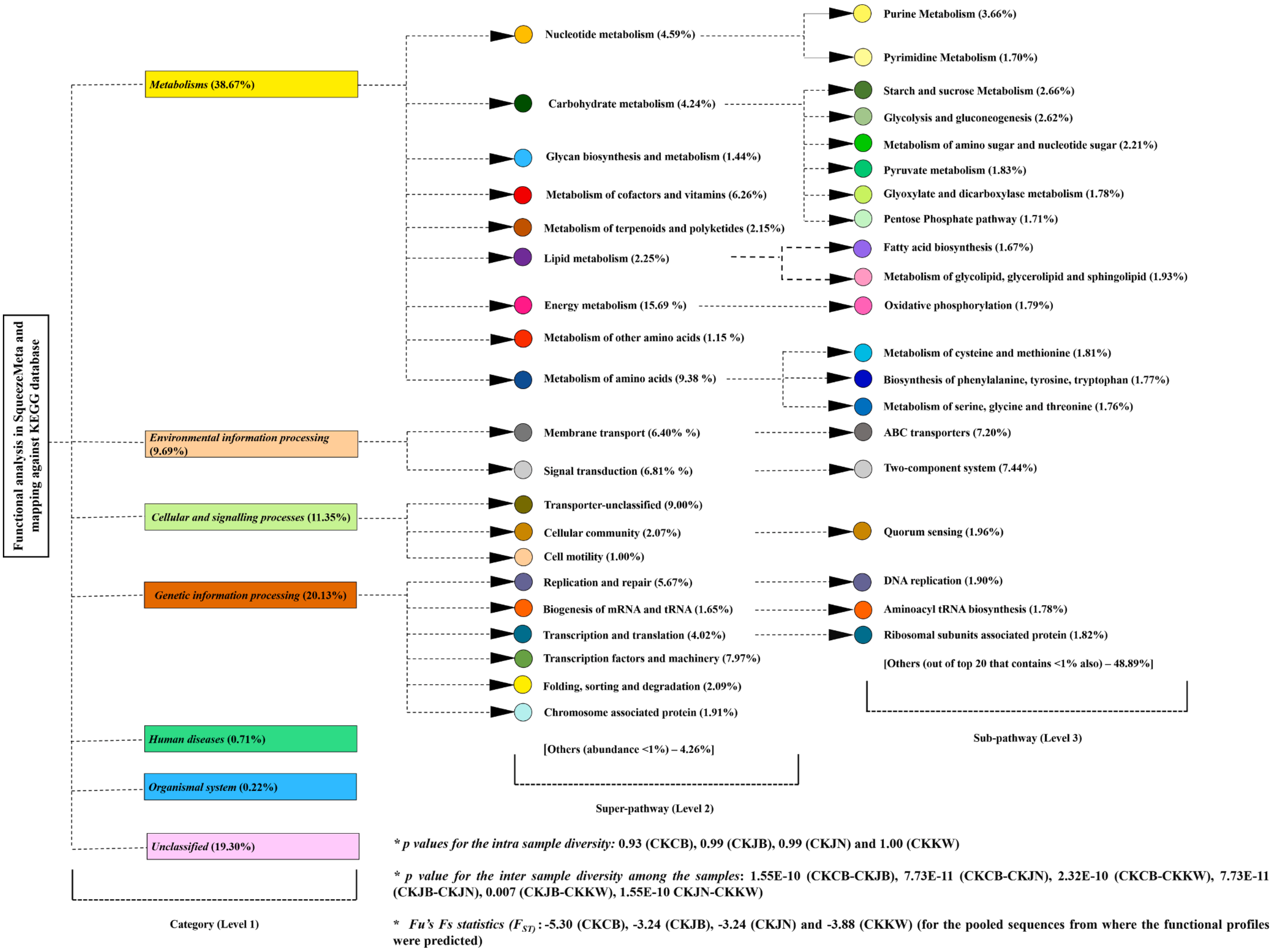
were found among the samples. Unique archaeal species in the samples of CKJB, CKJN and CKKW were 5, 3 and 2, respectively (Supplementary Table 12).

3.5. Functional profiles

Different enhanced functional pathways were observed after mapping metagenomic ORFs against eggNOG and KEGG databases. The number of prodigal ORFs identified from *chongkokjang* samples were 17,566 (CKCB), 40,397 (CKJB), 35,840 (CKJN) and 24,186 (CKKW), respectively. About 53 % of the total mapped ORFs were assigned to the COG functional genes and rest 47% were assigned to the KEGG functional pathways. In this study, we focussed more on the KEGG functional pathways rather than the output from COG mapping, due to output obtained from KEGG mapping was quite interesting. The functional pathways were categorized into three sub-classes via KEGG mapping. In level 1, metabolism was the most abundant category followed by genetic information processing, cellular and signalling processes, environmental information processing, human diseases, organismal systems and unclassified (Fig. 5). At level 2, 36 super-pathways and at level 3, 157 sub-pathways (terminal) from different categories were identified (Fig. 5) including some clustered regularly interspaced short palindromic repeats (CRISPR) associated proteins. Minor super-pathways and minor sub-pathways at < 1% abundance were also detected (Supplementary Tables 13 and 14).

Spearman's rank correlation was performed between some major microorganisms in *cheonggukjang* samples and different major level 2 super-pathways obtained from KO annotation. *Bacillus subtilis* and *B. thermoamylovorans* showed the significant positive correlation with the amino acid metabolism and metabolisms of vitamins and co-factors (Fig. 6a). *B. licheniformis* showed the significant positive correlation with the amino acid metabolisms and transporters. *Paenibacillus* sp., *Bacillus amyloliquifaciens*, *Bacillus stratosphericus* and *Bacillus velezensis* showed the significant positive correlation with biosynthesis of secondary metabolites. *Paenibacillus* sp. along with *Bacillus glycinifermentans* showed significant positive correlation to the nucleotide metabolism. *Bacillus coagulans*, *Carnobacterium inhibens*, *Terriglobus* sp. and *Bacillus andreraoultii* showed significant negative correlation with anti-microbial resistance (Fig. 6a). Based on the PCoA analysis, no clustering was found among the samples in terms of level 3 functional profiles (sub-pathways) (Fig. 6b).

One sample T test and Mann-whitney test were performed to check the intra sample and inter sample significance in terms of functional profiles of *cheonggukjang* metagenome. No intra sample significance was found in any of the pooled samples in terms of functional profiles. The p values for all 4 samples analysed from one sample T test were 0.93 (CKCB), 0.99 (CKJB), 0.99 (CKJN) and 1.00 (CKKW), respectively. However, based on Mann-whitney test, CKCB was significantly different from pooled CKJB (p = 1.55E-10), pooled CKJN (p = 7.73E-11) and pooled CKKW (p = 2.32E-10). Significant difference was found among CKJB-CKJN (p = 7.73E-11) and CKJB-CKKW (p = 0.007). Pooled CKKW was also significantly differed from pooled CKJN (p = 1.55E-10).



* *p values for the intra sample diversity:* 0.93 (CKCB), 0.99 (CKJB), 0.99 (CKJN) and 1.00 (CKKW)

* *p value for the inter sample diversity among the samples*: 1.55E-10 (CKCB-CKJB), 7.73E-11 (CKCB-CKJN), 2.32E-10 (CKCB-CKKW), 7.73E-11 (CKJB-CKJN), 0.007 (CKJB-CKKW), 1.55E-10 CKJN-CKKW)

* *Fu's Fs statistics* (F_{ST}) : -5.30 (CKCB), -3.24 (CKJB), -3.24 (CKJN) and -3.88 (CKKW) (for the pooled sequences from where the functional profiles were predicted)

Fig. 5. Different functional profiles (level 1, level 2 and level 3) obtained after mapping *cheonggukjang* metagenome against KEGG database. [*for the representation of level 2 and level 3 functional profiles, only top 20 pathways were considered based on mean abundance].

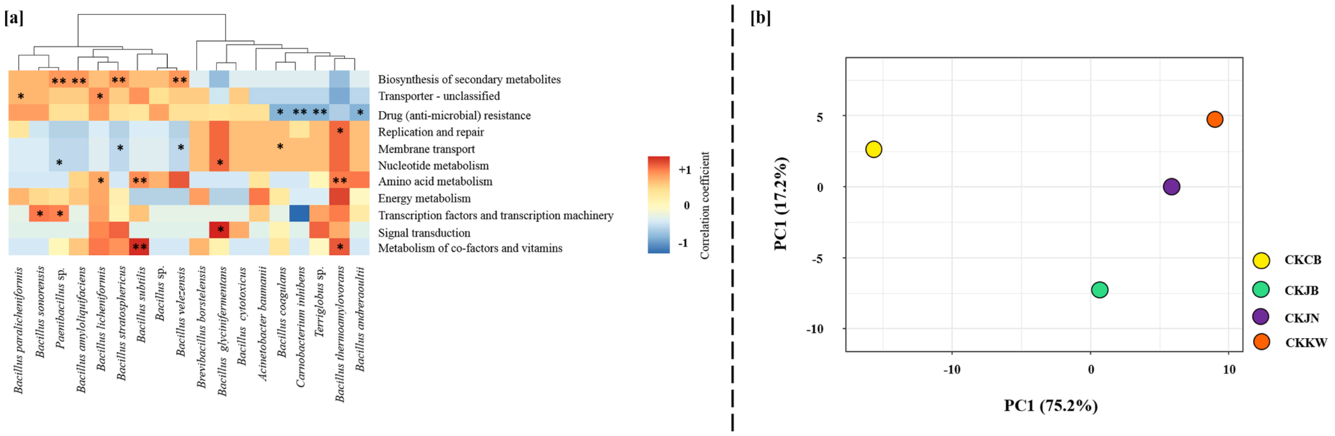
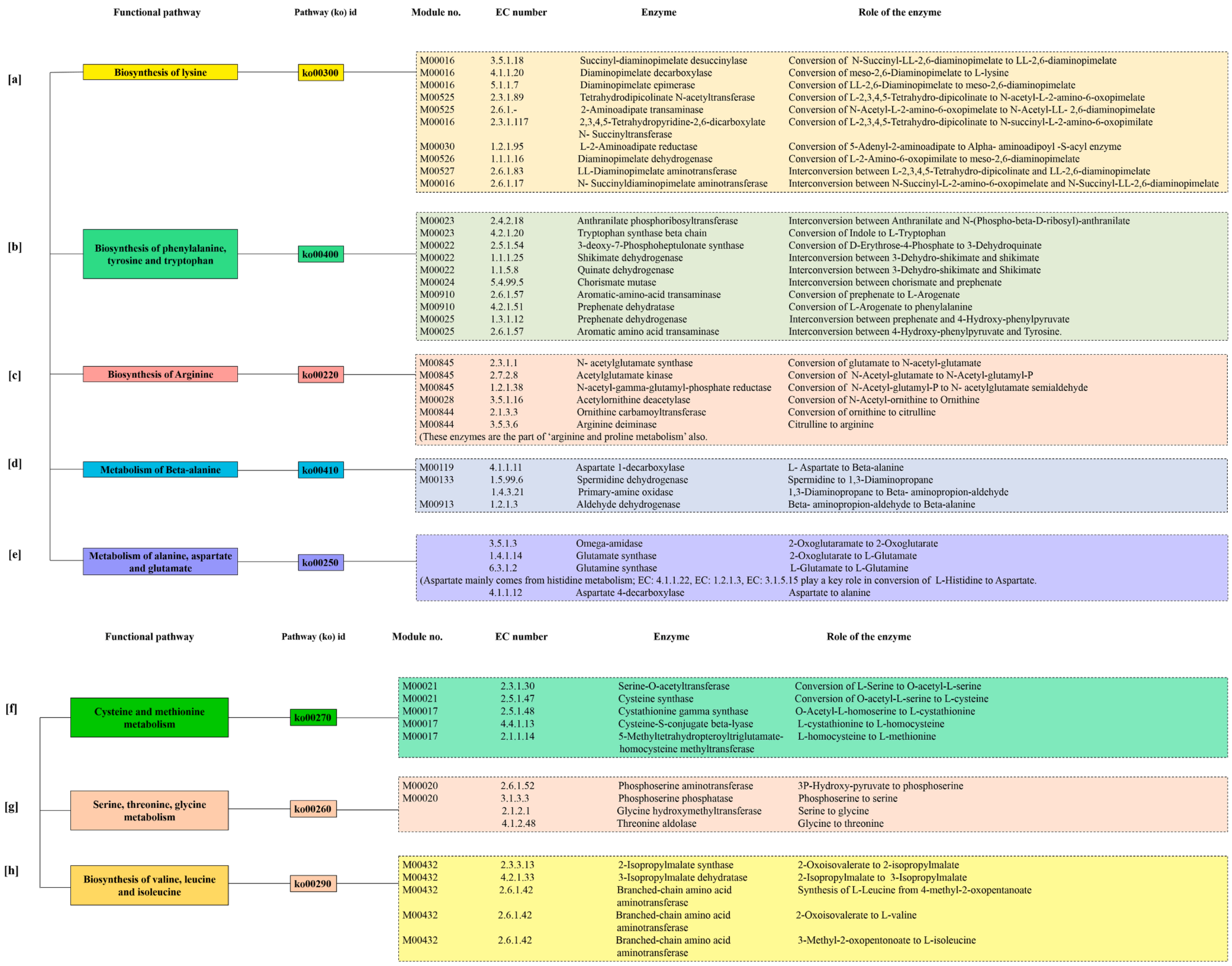


Fig. 6. (a) Non-parametric Spearman correlation among different level 2 super-pathways and the dominant species (identified from *cheonggukjang* metagenome) with mean abundance of >1%, (b) PCoA plot shoeing the difference among the samples in terms of level 3 sub-pathways.

Some predictive key enzyme for biosynthesis were detected such as diaminopimelate decarboxylase for lysine biosynthesis, tryptophan synthase beta chain for tryptophan biosynthesis, prephenate dehydratase for phenylalanine and aromatic amino acid transaminase for tyrosine biosynthesis, glutamate synthase and aspartate-4-decarboxylase for the synthesis of L-glutamate and alanine from 2-



	Functional pathway	Pathway (ko) id	Module no.	EC number	Enzyme	Role of the enzyme
[a]	Biosynthesis of lysine	ko00300	M00016	3.5.1.18	Succinyl-diaminopimelate desuccinylase	Conversion of N-Succinyl-LL-2,6-diaminopimelate to LL-2,6-diaminopimelate
			M00016	4.1.1.20	Diaminopimelate decarboxylase	Conversion of meso-2,6-Diaminopimelate to L-lysine
			M00016	5.1.1.7	Diaminopimelate epimerase	Conversion of LL-2,6-Diaminopimelate to meso-2,6-diaminopimelate
			M00525	2.3.1.89	Tetrahydrodipicolinate N-acetyltransferase	Conversion of L-2,3,4,5-Tetrahydro-dipicolinate to N-acetyl-L-2-amino-6-oxopimelate
			M00525	2.6.1.-	2-Aminoadipate transaminase	Conversion of N-Acetyl-L-2-amino-6-oxopimelate to N-Acetyl-LL- 2,6-diaminopimelate
			M00016	2.3.1.117	2,3,4,5-Tetrahydropyridine-2,6-dicarboxylate N- Succinyltransferase	Conversion of L-2,3,4,5-Tetrahydro-dipicolinate to N-succinyl-L-2-amino-6-oxopimilate
			M00030	1.2.1.95	L-2-Aminoadipate reductase	Conversion of 5-Adenyl-2-aminoadipate to Alpha- aminoadipoyl -S-acyl enzyme
			M00526	1.1.1.16	Diaminopimelate dehydrogenase	Conversion of L-2-Amino-6-oxopimilate to meso-2,6-diaminopimelate
			M00527	2.6.1.83	LL-Diaminopimelate aminotransferase	Interconversion between L-2,3,4,5-Tetrahydro-dipicolinate and LL-2,6-diaminopimelate
			M00016	2.6.1.17	N- Succinyldiaminopimelate aminotransferase	Interconversion between N-Succinyl-L-2-amino-6-oxopimelate and N-Succinyl-LL-2,6-diaminopimelate
[b]	Biosynthesis of phenylalanine, tyrosine and tryptophan	ko00400	M00023	2.4.2.18	Anthranilate phosphoribosyltransferase	Interconversion between Anthranilate and N-(Phospho-beta-D-ribosyl)-anthranilate
			M00023	4.2.1.20	Tryptophan synthase beta chain	Conversion of Indole to L-Tryptophan
			M00022	2.5.1.54	3-deoxy-7-Phosphoheptulonate synthase	Conversion of D-Erythrose-4-Phosphate to 3-Dehydroquinate
			M00022	1.1.1.25	Shikimate dehydrogenase	Interconversion between 3-Dehydro-shikimate and shikimate
			M00022	1.1.5.8	Quinate dehydrogenase	Interconversion between 3-Dehydro-shikimate and Shikimate
			M00024	5.4.99.5	Chorismate mutase	Interconversion between chorismate and prephenate
			M00910	2.6.1.57	Aromatic-amino-acid transaminase	Conversion of prephenate to L-Arogenate
			M00910	4.2.1.51	Prephenate dehydratase	Conversion of L-Arogenate to phenylalanine
			M00025	1.3.1.12	Prephenate dehydrogenase	Interconversion between prephenate and 4-Hydroxy-phenylpyruvate
			M00025	2.6.1.57	Aromatic amino acid transaminase	Interconversion between 4-Hydroxy-phenylpyruvate and Tyrosine.
[c]	Biosynthesis of Arginine	ko00220	M00845	2.3.1.1	N- acetylglutamate synthase	Conversion of glutamate to N-acetyl-glutamate
			M00845	2.7.2.8	Acetylglutamate kinase	Conversion of N-Acetyl-glutamate to N-Acetyl-glutamyl-P
			M00845	1.2.1.38	N-acetyl-gamma-glutamyl-phosphate reductase	Conversion of N-Acetyl-glutamyl-P to N- acetylglutamate semialdehyde
			M00028	3.5.1.16	Acetylornithine deacetylase	Conversion of N-Acetyl-ornithine to Ornithine
			M00844	2.1.3.3	Ornithine carbamoyltransferase	Conversion of ornithine to citrulline
			M00844	3.5.3.6	Arginine deiminase	Citrulline to arginine
			(These enzymes are the part of ‘arginine and proline metabolism’ also.			
[d]	Metabolism of Beta-alanine	ko00410	M00119	4.1.1.11	Aspartate 1-decarboxylase	L- Aspartate to Beta-alanine
			M00133	1.5.99.6	Spermidine dehydrogenase	Spermidine to 1,3-Diaminopropane
				1.4.3.21	Primary-amine oxidase	1,3-Diaminopropane to Beta- aminopropion-aldehyde
			M00913	1.2.1.3	Aldehyde dehydrogenase	Beta- aminopropion-aldehyde to Beta-alanine
[e]	Metabolism of alanine, aspartate and glutamate	ko00250		3.5.1.3	Omega-amidase	2-Oxoglutaramate to 2-Oxoglutarate
				1.4.1.14	Glutamate synthase	2-Oxoglutarate to L-Glutamate
				6.3.1.2	Glutamine synthase	L-Glutamate to L-Glutamine
			(Aspartate mainly comes from histidine metabolism; EC: 4.1.1.22, EC: 1.2.1.3, EC: 3.1.5.15 play a key role in conversion of L-Histidine to Aspartate.			
				4.1.1.12	Aspartate 4-decarboxylase	Aspartate to alanine

	Functional pathway	Pathway (ko) id	Module no.	EC number	Enzyme	Role of the enzyme
[f]	Cysteine and methionine metabolism	ko00270	M00021	2.3.1.30	Serine-O-acetyltransferase	Conversion of L-Serine to O-acetyl-L-serine
			M00021	2.5.1.47	Cysteine synthase	Conversion of O-acetyl-L-serine to L-cysteine
			M00017	2.5.1.48	Cystathionine gamma synthase	O-Acetyl-L-homoserine to L-cystathionine
			M00017	4.4.1.13	Cysteine-S-conjugate beta-lyase	L-cystathionine to L-homocysteine
			M00017	2.1.1.14	5-Methyltetrahydropteroyltriglutamate-homocysteine methyltransferase	L-homocysteine to L-methionine
[g]	Serine, threonine, glycine metabolism	ko00260	M00020	2.6.1.52	Phosphoserine aminotransferase	3P-Hydroxy-pyruvate to phosphoserine
			M00020	3.1.3.3	Phosphoserine phosphatase	Phosphoserine to serine
				2.1.2.1	Glycine hydroxymethyltransferase	Serine to glycine
				4.1.2.48	Threonine aldolase	Glycine to threonine
[h]	Biosynthesis of valine, leucine and isoleucine	ko00290	M00432	2.3.3.13	2-Isopropylmalate synthase	2-Oxoisovalerate to 2-isopropylmalate
			M00432	4.2.1.33	3-Isopropylmalate dehydratase	2-Isopropylmalate to 3-Isopropylmalate
			M00432	2.6.1.42	Branched-chain amino acid aminotransferase	Synthesis of L-Leucine from 4-methyl-2-oxopentanoate
			M00432	2.6.1.42	Branched-chain amino acid aminotransferase	2-Oxoisovalerate to L-valine
			M00432	2.6.1.42	Branched-chain amino acid aminotransferase	3-Methyl-2-oxopentonoate to L-isoleucine

Fig. 7. KO mapping of the various enzymes (with module number, EC number and functions) that were found to be involved in the synthesis of different amino acids identified in both metagenome and real time analysis: (a) biosynthesis of lysine, (b) biosynthesis of phenylalanine, arginine and tryptophan, (c) biosynthesis of arginine, (d) metabolism of beta-alanine, (e) metabolism of alanine, aspartate and glutamate, (f) cysteine and methionine metabolism, (g) serine, threonine and glycine metabolism, (h) biosynthesis of valine, leucine and isoleucine.

oxoglutarate and aspartate, respectively (Fig. 7). Branched chain amino acid aminotransferase was predicted for synthesis of leucine, isoleucine and valine (Fig. 7). Average relative viscosity of *cheonggukjang* samples were found within the range of 4–8 ± 1.43 cP to 0.5–1.0 ± 0.17 cP (Table 2). *Bacillus subtilis*, *B. sonorensis* and *B. amyloliquifaciens* showed the significant positive correlation with viscosity (Supplementary Table 15).

3.6. Validation of KO annotated amino acid profiles through experimentally analysed free amino acids

Amino acid profiles obtained from KEGG annotation in *cheonggukjang* samples were validated with amino acid profiles analysed by amino acid analyser to study the real time presence of free amino acids in the food matrix (Table 3). Some KO annotated predicted pathways and related essential amino acids (Supplementary Table 16) viz. tyrosine, phenylalanine, valine, leucine, isoleucine, beta-alanine, glycine, serine, threonine, GABA and cysteine were validated during this study, as all these amino acid were found to be detected through amino acid analyser (Table 3 and Supplementary Fig. 1). Some amino acids (sarcosine, ornithine, asparagine) were detected in real time analysis but they were not detected in predictive analysis (Table 3 and Supplementary Fig. 1). Contrastingly, methionine, tryptophan, glutathione, taurine and hypotaurine were detected in predictive analysis only (Table 3 and Supplementary Fig. 1). Clear difference was observed between the predictive data and real time data of amino acid profile; overall, the abundance of the amino acids predicted computationally was found lower than the amino acids detected by amino acid analyser. In predictive analysis, highest abundance was observed for the prediction of arginine biosynthesis followed by the prediction for proline, serine and phenylalanine; however, glycine was found to be the most dominant in real time data followed by glutamic acid, histidine and tyrosine (Table 3).

4. Discussion

4.1. Metataxonomic analysis

Cheonggukjang is considered as one of the oldest fermented soybean foods in Korean gastronomy (Kwon, Chung, & Jang, 2019), and is traditionally prepared by natural fermentation (Shin, Kim, Park, & Kim, 2016). Nowadays, *cheonggukjang* is commonly produced in Korea by using commercial strains of *Bacillus* spp. (Lee et al., 2015). Profiling of microbial community in naturally fermented *cheonggukjang* by shotgun metagenomic method has not reported yet. Hence, we applied shotgun metagenomic method to profile the whole microbial community structure in naturally fermented *cheonggukjang.* In our study, we collected each sample of naturally fermented *cheonggukjang* from 10 different regions of Republic of Korea viz. Gangjin-gundong *cheonggukjang* from Gangjin (Jeonnam province), Sam-gol *cheonggukjang*, Urikon-umni *cheonggukjang*, Duseungsan-kongmeaul cheonggukjang, Myeongwoldank *cheonggukjang* from Jeongup (Jeonbuk province), Hanga gol *cheonggukjang* from Chungju (Chungbuk) and Gotoil cheonggukjang from Taebaekk (Kangwon province). Ultimately, 4 samples were sequenced (*pool-seq*) from representative pooled DNA (based on the Province) for shotgun sequence analysis, since *pool seq* can also represent the actual sequences from the pooled DNA from each sample (Anand et al., 2016) to maintain the adequate diversity within and between the samples. As these samples are naturally fermented, our expectation was to observe the variation of microbial community structure within a single pool representative. Statistics were applied to interpret the *pool seq* on the basis of nucleotide differences and neutrality test, and clearly indicated the low polymorphism and variation within the pooled sample which signifies that the samples from the same origin possess the minimal diversity differences (Hivert, Leblois, Petit, Gautier, & Vitalis, 2018). Negative F_{ST} values are also clear evidence for an excess number of alleles in a single pooled sample, which actually represents more than one sample (De Meeûs, 2018). So, it is clear from these two tests that each of our pooled sample actually represented the combination of more than one sample and there is very less intra-sample diversity within the pooled sample, which can also justify

Table 3
Comparison and validation of KO predicted amino acids with the real time data obtained from amino acid analyser.

Name of the amino acids	Amino acids detected by amino acid analyser				Predicated amino acid inferred by KO annotation			
	Relative abundance (%) (Mean ± SD)				Relative abundance (%)			
	CKCB	CKJB	CKJN	CKKW	CKCB	CKJB	CKJN	CKKW
Lysine	1.36 ± 0.62	2.19 ± 0.81	2.01 ± 1.07	1.82 ± 0.35	0.68	0.73	0.75	0.78
Beta-alanine	5.26 ± 1.06	7.32 ± 0.06	7.0 ± 1.44	5.90 ± 1.31	0.20	0.18	0.19	0.20
Taurine and hypotaurine	Not detected from experimental analysis				0.28	0.28	0.27	0.28
Methionine	Not detected from experimental analysis				1.19	1.80	1.84	1.82
Cysteine	2.26 ± 3.19	4.82 ± 0.33	5.40 ± 0.46	4.20 ± 0.7				
Sarcosine	0	0.11 ± 0.19	0.08 ± 0.14	0.27 ± 0.08	Not detected from predictive analysis			
Ornithine	2.79 ± 0.29	3.70 ± 0.73	2.84 ± 0.34	3.00 ± 0.06	Not detected from predictive analysis			
Serine	0.89 ± 0.27	0.94 ± 0.30	0.81 ± 0.33	0.90 ± 0.37	1.76	1.75	1.76	1.75
Glycine	12.02 ± 1.51	11.31 ± 1.90	10.82 ± 1.04	10.51 ± 0.43				
Threonine	1.74 ± 0.54	1.86 ± 0.24	1.85 ± 0.37	1.81 ± 0.76				
Arginine	2.11 ± 0.01	1.00 ± 0.88	1.78 ± 1.00	0.82 ± 1.16	2.04	1.92	1.86	1.85
Proline	0	0	0	0	2.00	1.93	1.83	1.81
GABA	3.42 ± 2.79	6.94 ± 2.59	5.83 ± 1.17	8.58 ± 3.19	Not detected from predictive analysis			
Histidine	5.79 ± 0.89	8.84 ± 0.48	7.64 ± 1.08	5.81 ± 0.80	0.45	0.44	0.44	0.44
Glutathione	Not detected from experimental analysis				0.33	0.33	0.34	0.32
Phenylalanine	5.11 ± 0.82	3.79 ± 3.28	4.90 ± 1.18	4.71 ± 1.21	1.75	1.76	1.78	1.78
Tyrosine	5.69 ± 2.02	7.63 ± 0.74	8.08 ± 1.11	5.81 ± 0.80				
Tryptophan	Not detected from experimental analysis							
Valine	2.22 ± 3.14	0	0	0	0.50	0.51	0.50	0.50
Leucine	3.81 ± 3.3	1.17 ± 0.14	1.45 ± 0.34	1.01 ± 0.2				
Isoleucine	2.58 ± 2.44	5.87 ± 0.21	5.48 ± 1.70	4.93 ± 0.86				
Alanine	0.92 ± 0.08	1.15 ± 0.23	0.97 ± 0.39	1.18 ± 0.19	1.51	1.56	1.56	1.57
Aspartic acid	2.58 ± 1.74	2.35 ± 0.47	3.74 ± 0.75	3.45 ± 1.74				
Glutamic acid	10.90 ± 1.50	11.40 ± 0.01	13.35 ± 2.10	15.22 ± 5.13				
Asparagine	2.45 ± 1.35	1.33 ± 1.20	2.17 ± 0.7	2.02 ± 1.58	Not detected from predictive analysis			
Beta-AiBA	4.98 ± 5.31	2.63 ± 1.25	2.71 ± 0.81	3.00 ± 0.49	Not detected from predictive analysis			

our pooling.

Firmicutes was the most abundant phylum in *cheonggukjang*, like in other Asian fermented soybean foods such as *kinema* (Kumar et al. 2019; Kharnaior, & Tamang, 2021), *douchi* of China (Yang, Yang, Tu, & Wang, 2016), and *doenjang* (Jung, Jung, Lee, & Jeon, 2016). *Bacillaceae* was the abundant family in *cheonggukjang* as reported earlier in other ethnic soybean fermented foods (Jung, Jung, Lee & Jeon, 2016; Kim, Kwak, Jung, & Kim, 2016; Nam, Yi, & Lim, 2012). A colossal diversity of *Bacillus* with 150 species was observed in *cheonggokjang*, showing inter-species diversity (Dabire, Somda, Ugwuanyi, Ezeogu, & Traore, 2018). Inter-species diversity of *Bacillus* was also reported in *kinema* (Tamang, Jeyaram, Rai, & Mukherjee, 2021), *doenjang* (Jung, Jung, Lee & Jeon, 2016; Kim, Kwak, Jung, & Kim, 2016), *dawadawa* of Africa (Akanni, Naudé, De Kock, & Buys, 2018). *Bacillus thermoamylovorans* was the most abundant bacterial species in *cheonggukjang* followed by *Bacillus licheniformis*, *Bacillus glycinifermentans*, *Bacillus subtilis* and *Bacillus paralicheniformis*. *B. thermoamylovorans* is a heat resistant bacterium (Berendsen et al., 2015), which is reported earlier in *cheonggukjang* (Lee et al., 2016), *kinema* (Kharnaior, & Tamang, 2021) and *douchi* (Chen et al., 2011). The shotgun sequence analysis of *cheonggokjang* samples revealed the dominance of *Bacillus thermoamylovorans* as the most abundant species in contrast to earlier report of *Bacillus subtilis* as the most dominant species in *cheonggokjang* analysed by pyrosequencing method (Nam, Yi, & Lim, 2012) and cultural methods (Kindoli, Lee, Heo, & Kim, 2012; Moon et al., 2015). *B. thermoamylovorans* is the thermotolerant and amylolytic bacterium (Choonut, Prasertsan, Klomklao, & Sangkharak, 2020). *B. subtilis* is one of the major bacterial species in many Asian fermented soybean foods (Kubo et al., 2011; Chettri, & Tamang, 2015; Tamang, 2015; Ju et al., 2019; Kharnaior, & Tamang, 2021). *B. licheniformis* is known to be a probiotic organism for human consumption (Lee, Kim, & Paik, 2019) and also produces extracellular nuclease, *Nuc B* (Baslé et al., 2018). Presence of *B. paralicheniformis* was also reported in *cheonggukjang* (Dunlap, Kwon, Rooney, & Kim, 2015). Though in a minor abundance, detection of amino acid-producing bacteria *Corynebacterium glutami*cum (Hahne, Kloster, Rathmann, Weber, & Lipski, 2018) and *B. thuringiensis* (Hibi et al., 2011) in *cheonggokjang* indicated the nutritive significance of *cheonggukjang*.

Though, the cumulative abundance of lactic acid bacterial species was only 0.8%, however, presence of *Lactobacillus*, *Enterococcus*, *Lactococcus*, *Aerococcus*, *Carnobacterium*, *Streptococcus*, *Vagococcus*, *Tetragenococcus* and *Weissella* were detected in *cheonggukjang* samples. After bacteria, virus was the most abundant domain found in the metagenome analysis of *cheonggukjang*. The presence of *Myoviridae, Siphoviridae, Podoviridae* and *Mimiviridae* supports the previous findings from the studies on *kimchi*, fermented *shrimp* and *sauerkraut* (Park et al., 2011). *Bacillus* phages were found dominant over other phages in *cheonggukjang* metagenome. Prevalence of *Bacillus* phages (mostly against *B. subtilis*) was also revealed from the earlier study on *cheonggukjang* (Ghosh, Kang, Hyun, & Kim, 2018). Bacteriophages generally play the critical role in determining the bacterial abundance over the time (Shapiro, Kushmaro, & Brenner, 2010). It is quite obvious that the higher abundance of bacteriophages related to the key host, i.e. *Bacillus* phages in *cheonggukjang* fermentation, can be correlated with the deterioration of the product quality (Ghosh, Kang, Hyun, & Kim, 2018; Lee, 1978); but in our study, the abundance of *Bacillus* phages was found much lower than the abundance of *Bacillus* species. However, the presence of phages against the hosts like *Staphylococcus*, *Clostridium* and *Escherichia*, indicates the bio-control activity of bacteriophages in food fermentation (Kumar et al., 2019).

Nevertheless, any direct interaction of bacteria with the phage was not found in the *cheonggukjang* metagenome but, we found some CRISPR associated proteins from the predictive analysis of *cheonggukjang* metagenome, which were already reported as one of the anti-phage defence mechanisms of bacteria and archaea (Rostøl, & Marraffini, 2019). From the predictive analysis of *cheonggukjang* metagenome, we have identified Cas3 (type I system), Cas 9 (type II system), different Csm (csm 6) and Cmr (Cmr 4,5,6) proteins (part of type III system), CasD proteins, Cst proteins and Csd proteins. Type I and type II systems were reported to destroy their invaders using the crRNA guided inactivation of the protospacer sequence in the genome of the phage (Gasiunas et al., 2012); whereas, type III system is known to target the transcription machinery of the invaders, which cuts the non-template strand of the transcribed viral DNA (Samai et al., 2015). Type II is also known to employ a single-subunit RNA-guided nuclease (Cas9) to combat the viruses (phage) (Sapranauskas et al., 2011). No prominent feature of phage for escaping CRISPR defence was found in our study. We can match up this functional feature to the CRISPR based anti-phage defence mechanism of bacteria against the virus; earlier we also mentioned the fact that, phage abundance was found much lower than the abundance of bacterial community in *cheonggukjang*. Moreover, we may assume this phenomenon as the key factor for the higher abundance of fermenting bacteria over phage abundance and infection that determines the quality of the product too. Still we believe that, further investigation is also needed to decipher the viral-bacterial interaction more firmly.

Few species of haloarchaea were detected in *cheonggukjang*, though their abundance was found very low. Exact role of archaea in *cheonggukjang* is still unknown. Halophilic archaea has been reported as the producer of halocin, a compound that infers the anti-microbial activity (Corral, Amoozegar, & Ventosa, 2019). Hence the presence of haloarchaea again may be linked with the bio-enhancement of the fermented soybean food (Lee, 2013). Domain Eukarya was which consisted of yeasts, filamentous moulds, different species of algae, protozoa and parasites was detected in low abundances in *cheonggukjang*. Other micro-eukaryotes, including protozoa and parasites are thought to possess the negative impact on the health and well-being of the consumers; though complete safety evaluation regarding the fermented soybean food is yet to be studied. Shotgun metagenome sequence is a powerful tool which can detect an entire microbial community in sample with high and low abundances (Quince, Walker, Simpson, Loman, & Segata, 2017), this may help to understand the safety measure of foods on the basis of their abundances level (Grützke et al., 2019).

4.2. *Functional profiles*

The KEGG annotation revealed the prevalence of metabolism in *cheonggukjang*. Amino acid metabolism was found as one of the dominant predictive functional super-pathways under the category of metabolisms. Metabolism and biosynthesis of alanine, aspartate and glutamate enhance the production of flavour and aroma compound during fermentation (Sulaimam, Gan, Yin, & Chan, 2014; An et al., 2020). Glutamic acid forms the poly-glutamic acid (PGA), a polymer of the corresponding amino acid (Hsueh, Huang, Kunene, & Lee, 2017), which imparts the characteristics the stickiness of the fermented soybean food (Chettri, Bhutia, & Tamang, 2016). Abundance of the genes involved in the synthesis of branched chain amino acids (valine, leucine and isoleucine) were also detected from the metagenome analysis of *cheonggukjang*; which involve in energy metabolisms in *Bacillus* spp. (Belitsky, 2015). Amino acid profiles of *cheonggukjang* also validated the predictive amino acid profiles as inferred by KEGG annotation. Glycine, alanine and lysine are known to play role in formation of aroma and taste in fermented soybean foods (Zhao, Schieber, & Gänzle, 2016). Prediction of the synthesis and metabolism of glutathione was inferred from the functional analysis of *cheonggukjang* metagenome. Metabolism, biosynthesis and transport of glutathione participate in many cellular events viz. anti-oxidant defence, drug detoxification and cell signalling involved in the regulation of gene expression and apoptosis (Kennedy, Sandhu, Harper, & Cuperlovic-Culf, 2020). Presence of the gene involved in the synthesis of taurine and hypotaurine were also predicted computationally, which have been reported to confer the health benefits against cardiovascular diseases (Xu, Arneja, Tappia, & Dhalla, 2008). Moreover, the prediction of different amino acids were validated with the real time analysis of amino acids. A wide range of essential amino acids predicted through KO annotation were also found in the real time analysis which confirms the presence of those predicted amino acids in the food matrix. Some pathways of amino acid biosynthesis were obtained from KEGG annotation only (methionine, tryptophan, taurine and hypotaurine) but were not detected in real time analysis, may be due to the repression of corresponding genes in the microbes during the fermentation (Zhi, Wu, & Xu, 2017). The overall abundance of the predictive amino acid profile was found quite lower than the amino acid profile obtained from real time analysis; this may be corresponded to some strategical disadvantages of metagenome functionality prediction (Prakash, & Taylor, 2012). While using the metagenomic contigs, sometimes it remains difficult to understand the proper microbial niche/environment which ultimately determines the functional aspect of microbial communities and degree of various gene expression (Prakash, & Taylor, 2012). Additionally, attributes of metagenomic sequence such as sequence depth, length of query sequences, sequencing coverage and the specific source of sequences of any particular functional features were already reported to affect the prediction of functional features or any functional gene expression (Prakash, & Taylor, 2012). Though, more studies on genes and gene expression are needed to establish this fact firmly.

Prediction of biosynthesis of folate, vitamin B7 (biotin), vitamin A (retinol) and riboflavin was observed in *cheonggukjang*. Folate, biotin and riboflavin are produced during soybean fermentation (Jayachandran, & Xu, 2019). Though prediction of some pathways related to human disease were also observed, but their abundance were too low to make any significant impact. Moreover, the positive attributes were found to overrule the negative impact of those pathways that have the potency to cause any harm.

This present study may also provide a theoretical and technical support for searching and applying the key enzymes, bio-active compounds and different essential amino acids with health benefits, which were predicted during *cheonggukjang* fermentation for other Asian fermented soybean foods. The metataxonomic and predictive functional features of the *cheonggukjang* metagenome may be used to design the starter culture with enhanced nutritional quality under the controlled and hygienic fermentation condition.

5. Conclusion

Cheonggokjang is a delicacy of Korean diets and presents one of the ancient foods of Korea. Several researchers have studied microbiology and nutritional aspects of *cheonggokjang*. Our study revealed different beneficial microorganisms associated with *cheonggokjang* fermentation along with their beneficial predictive functional profiles. We hope that the information obtained from this study may help to sensitise the commercial producers and consumers aware on microbial community the health benefits, hygiene and general safety in *cheonggokjang*. We believe this is the first report on shotgun-based metataxonomic profile of naturally fermented *cheonggukjang*.

6. Data availability

The metagenomic sequences generated in the present study are available in the NCBI database with the SRA accession numbers SRR13529508 (CKCB), SRR13529507 (CKJB), SRR13529506 (CKJN) and SRR13529505 (CKKW) under the bioproject id PRJNA694851.

CRediT authorship contribution statement

Jyoti Prakash Tamang: Conceptualization, Funding acquisition, Investigation, Project administration, Software, Supervision, Validation, Visualization, Writing – original draft. **Souvik Das:** Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft. **Pynhunlang Kharnaior:** Formal analysis. **Priyambada Pariyar:** Formal analysis. **Namrata Thapa:** Conceptualization, Investigation, Resources, Supervision. **Seung-Wha Jo:** Data curation, Formal analysis, Methodology, Validation. **Eun-Jung Yim:** Data curation, Formal analysis, Methodology, Validation. **Dong-Hwa Shin:** Conceptualization, Supervision, Visualization, Writing – original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

JPT, SD, PK, PP and NT are grateful to the Department of Biotechnology (DBT), Ministry of Science and Technology, Government of India for financial support through DBT-DAICENTER project, sanction number: BT/B1/14/042/2017. SWJ, EJY and DHS are grateful to Shindonghwa Food Research Institute, Seoul and Microbial Institute for Fermentation Industry(MIFI), Sunchang, respectively for the support.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodres.2021.110904.

References

Akanni, G. B., Naudé, Y., De Kock, H. L., & Buys, E. M. (2018). Diversity and functionality of bacillus species associated with alkaline fermentation of bambara groundnut (*Vigna subterranean* L. Verdc) into dawadawa-type African condiment. *European Food Research and Technology, 244*. https://doi.org/10.1007/s00217-017-3024-x

Ali, M. W., Shahzad, R., Bilal, S., Adhikari, B., Kim, I. D., Lee, J. D., … Shin, D. H. (2018). Comparison of antioxidants potential, metabolites, and nutritional profiles of Korean fermented soybean (*Cheonggukjang*) with *Bacillus subtilis* KCTC 13241. *Journal of Food Science and Technology, 55*(8), 2871–2880.

An, F., Li, M., Zhao, Y., Zhang, Y., Mu, D., et al. (2020). Metatranscriptome-based investigation of flavor-producing core microbiota in different fermentation stages of dajiang, a traditional fermented soybean paste of Northeast China. *Food Chemistry, 1* (343), 128509. https://doi.org/10.1016/j.foodchem.2020.128509

Anand, S., Mangano, E., Barizzone, N., Bordoni, R., Sorosina, M., Clarelli, F., … De Bellis, G. (2016). Next generation sequencing of pooled samples: Guideline for variants' filtering. *Scientific Reports, 6*, 33735. https://doi.org/10.1038/srep33735

Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., & Zhang, L. (2018). DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome, 6*, 23. https://doi.org/10.1186/s40168-018-0401-z

Arıkan, M., Mitchell, A. L., Finn, R. D., & Gürel, F. (2020). Microbial composition of Kombucha determined using amplicon sequencing and shotgun metagenomics. *Journal of Food Science, 85*, 455–464.

Bae, M. J., Shin, H. S., See, H. J., Chai, O. H., & Shon, D. H. (2014). Cheonggukjang ethanol extracts inhibit a murine allergic asthma via suppression of mast cell-dependent anaphylactic reactions. *Journal of Medicinal Food, 17*(1), 142–149.

Baek, J. G., Shim, S. M., Kwon, D. Y., Choi, H. K., Lee, C. H., & Kim, Y. S. (2010). Metabolite profiling of Cheonggukjang, a fermented soybean paste, inoculated with various *Bacillus* strains during fermentation. *Bioscience, Biotechnology, Biochemistry, 74*(9), 1860–1868.

Bang, M. S., Jeong, H. W., Lee, Y. J., Lee, S. C., Lee, G. S., Kim, S., et al. (2020). Complete genome sequence of Bacillus velezensis strain DKU_NT_04, isolated from a traditional Korean food made from soybeans (cheonggukjang). *Microbiology Resource Announcements, 9*(24), e00477–20. https://doi.org/10.1128/MRA.00477-20

Baslé, A., Hewitt, L., Koh, A., Lamb, H. K., Thompson, P., Burgess, J. G., … Lewis, R. J. (2018). Crystal structure of NucB, a biofilm-degrading endonuclease. *Nucleic Acids Research, 46*(1), 473–484.

Belitsky, B. R. (2015). Role of branched-chain amino acid transport in *Bacillus subtilis* CodY activity. *Journal of Bacteriology, 197*(8), 1330–1338.

Berendsen, E. M., Krawczyk, A. O., Klaus, V., de Jong, A., Boekhorst, J., Eijlander, R. T., … Wells-Bennik, M. H. J. (2015). *Bacillus thermoamylovorans* spores with very-high-level heat resistance germinate poorly in rich medium despite the presence of *ger* clusters but efficiently upon exposure to calcium-dipicolinic acid. *Applied and Environmental Microbiology, 22*, 7791–7801.

Buchfink, B., Xie, C., & Hutson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods, 12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Cao, Y., Fanning, S., Proos, S., Jordan, K., & Srikumar, S. (2017). A review on the applications of next generation sequencing technologies as applied to food-related microbiome studies. *Frontiers in Microbiology, 8*, 1829. https://doi.org/10.3389/fmicb.2017.01829

Chen, C., Huang, H., & Wu, C. H. (2017). Protein bioinformatics databases and resources. *Methods in Molecular Biology (Clifton. N.J.), 1558*, 3–39.

Chen, T., Xiong, S., Jiang, S., Wang, M., Wu, Q., & Wei, H. (2011). Molecular identification of microbial community in Chinese douchi during post-fermentation process. *Food Science and Biotechnology, 20*, 1633–1638.

Chettri, R., & Tamang, J. P. (2015). *Bacillus* species isolated from tungrymbai and bekang, naturally fermented soybean foods of India. *International Journal of Food Microbiology, 197*, 72–76.

Chettri, R., Bhutia, M. O., & Tamang, J. P. (2016). Poly-γ-glutamic acid (PGA)-producing *Bacillus* species isolated from Kinema, Indian fermented soybean food. *Frontiers in Microbiology, 7*, 971. https://doi.org/10.3389/fmicb.2016.00971

Choonut, A., Prasertsan, P., Klomklao, S., & Sangkharak, K. (2020). Study on mcl-PHA Production by Novel Thermotolerant Gram-Positive Isolate. *Journal of Polymers and the Environment, 28*, 2410–2421.

Corral, P., Amoozegar, M. A., & Ventosa, A. (2019). Halophiles and their biomolecules: Recent advances and future applications in biomedicine. *Marine Drugs, 18*(1), 33. https://doi.org/10.3390/md18010033

Dabire, Y., Somda, M. K., Ugwuanyi, J., Ezeogu, L. I., & Traore, A. S. (2018). Diversity of *Bacillus* species and their antimicrobial compounds involved in alkaline-fermentation of indigenous food condiments used in Africa. *International Journal of Advanced Research, 6*, 331–355.

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., & Van Broeckhoven, C. (2018). NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics, 34* (15), 2666–2669.

De Meeûs, T. (2018). Revisiting F_{IS}, $_{FS}$T, wahlund effects, and null alleles. *Journal of Heredity, 109*(4), 446–456.

Dunlap, C. A., Kwon, S. W., Rooney, A. P., & Kim, S. J. (2015). *Bacillus paralicheniformis* sp. nov., isolated from fermented soybean paste. *International Journal of Sysstematic and Evolutionary Microbiology, 65*(10), 3487–3492.

Galperin, M. Y., Kristensen, D. M., Makarova, K. S., Wolf, Y. I., & Koonin, E. V. (2019). Microbial genome analysis: The COG approach. *Briefings in Bioinformatics, 20*(4), 1063–1070.

Gasiunas, G., Barrangou, R., Horvath, P., & Siksnys, V. (2012). Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences, 109*(39), E2579–E2586. https://doi.org/10.1073/pnas.1208507109

Ghosh, K., Kang, H. S., Hyun, W. B., & Kim, K. P. (2018). High prevalence of *Bacillus subtilis*-infecting bacteriophages in soybean-based fermented foods and its detrimental effects on the process and quality of Cheonggukjang. *Food Microbiology, 76*, 196–203.

Grützke, J., Malorny, B., Hammerl, J. A., Busch, A., Tausch, S. H., Tomaso, H., & Deneke, C. (2019). Fishing in the soup – pathogen detection in food safety using metabarcoding and metagenomic sequencing. *Frontiers in Microbiology, 10*, 1805. https://doi.org/10.3389/fmicb.2019.01805

Hahne, J., Kloster, T., Rathmann, S., Weber, M., & Lipski, A. (2018). Isolation and characterization of *Corynebacterium* spp. from bulk tank raw cow's milk of different dairy farms in Germany. *PLoS ONE, 13*(4), e0194365. https://doi.org/10.1371/journal.pone.0194365

Hartley, I. E., Liem, D. G., & Keast, R. (2019). Umami as an 'alimentary' taste. a new perspective on taste classification. *Nutrients, 11*(1), 182. https://doi.org/10.3390/nu11010182

Hibi, M., Kawashima, T., Kodera, T., Smirnov, S. V., Sokolov, P. M., Sugiyama, M., … Ogawa, J. (2011). Characterization of *Bacillus thuringiensis* L-isoleucine dioxygenase for production of useful amino acids. *Applied and Environmental Microbiology, 77*(19), 6926–6930.

Hivert, V., Leblois, R., Petit, E. J., Gautier, M., & Vitalis, R. (2018). Measuring genetic differentiation from pool-seq data. *Genetics, 210*(1), 315–330.

Hong, S. W., Choi, J. Y., & Chung, K. S. (2012). Culture-based and denaturing gradient gel electrophoresis analysis of the bacterial community from *Chungkookjang*, a traditional Korean fermented soybean food. *Journal of Food Science, 77*(10), M572–M578.

Hong, S. W., Lim, I., Kim, Y., Shin, S. M., & Chung, K. (2013). Denaturing gradient gel electrophoresis and culture-based analysis of the bacterial community in cheonggukjang, a Korean traditional fermented soybean food from Gangwon province. *Korean Journal of Food Science and Technology, 45*. https://doi.org/10.9721.KJFST.2013.45.4.515.

Hsueh, Y. H., Huang, K. Y., Kunene, S., & Lee, T. Y. (2017). Poly-γ-glutamic acid synthesis, gene regulation, phylogenetic relationships, and role in fermentation. *International Journal of Molecular Sciences, 18*(12), 2644. https://doi.org/10.3390/ijms18122644

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., … Jensen, L. J. (2016). eggNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research, 44*(D1), D286–D293.

Hyatt, D., LoCascio, P. F., Hauser, L. J., & Uberbacher, E. C. (2012). Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics, 28*(17), 2223–2230.

Jayachandran, M., & Xu, B. (2019). An insight into the health benefits of fermented soy products. *Food Chemistry, 271*, 362–371.

Ju, S., Cao, Z., Wong, C., Liu, Y., Foda, M. F., Zhang, Z., & Li, J. (2019). Isolation and optimal fermentation condition of the *Bacillus subtilis* subsp. natto strain WTC016 for nattokinase production. *Fermentation, 5*, 92. https://doi.org/10.3390/fermentation5040092

Jung, W. Y., Jung, J. Y., Lee, H. J., & Jeon, C. O. (2016). Functional characterization of bacterial communities responsible for fermentation of Doenjang: A traditional Korean fermented soybean paste. *Frontiers in Microbiology, 7*, 827. https://doi.org/10.3389/fmicb.2016.00827

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research, 28*(1), 27–30.

Kennedy, L., Sandhu, J. K., Harper, M. E., & Cuperlovic-Culf, M. (2020). Role of glutathione in cancer: From mechanisms to therapies. *Biomolecules, 10*, 1429. https://doi.org/10.3390/biom10101429

Kharnaior, P., & Tamang, J. P. (2021). Bacterial and fungal communities and their predictive functional profiles in *kinema*, a naturally fermented soybean food of India, Nepal and Bhutan. *Food Research International, 140*, 110055. https://doi.org/10.1016/j.foodres.2020.110055

Kim, J., Choi, J. M., Maria John, K. M., Kusano, M., Oikawa, A., Saito, K., & Lee, C. H. (2012). GC–TOF-MS- and CE–TOF-MS-based metabolic profiling of *cheonggukjang* (fast-fermented bean paste) during fermentation and its correlation with metabolic pathways. *Journal of Agricultural and Food Chemistry, 60*(38), 9746–9753.

Kim, J., Choi, J. N., Kang, D., Son, G. H., Kim, Y. S., Choi, H. K., … Lee, C. H. (2011). Correlation between antioxidative activities and metabolite changes during Cheonggukjang fermentation. *Bioscience, Biotechnology, Biochemistry, 75*(4), 732–739.

Kim, J. W., Kim, Y. S., Jeong, P. H., Kim, H. E., & Shin, D. H. (2006). Physicochemical characteristics of traditional fermented soybean products manufactured in folk villages of Sunchang region. *Journal of Food Hygiene and Safety, 21*, 223–230.

Kim, M. J., Kwak, H. S., Jung, H. Y., & Kim, S. S. (2016). Microbial communities related to sensory attributes in Korean fermented soybean paste (*doenjang*). *Food Research International, 89*, 724–732.

Kim, M. J., Lee, D. H., Ahn, J., Jang, Y. J., Ha, T. Y., Do, E., & Jung, C. H. (2020). Nutri-kinetic study of fermented soybean paste (*Cheonggukjang*) isoflavones according to the Sasang typology. *Nutrition Research and Practice, 14*(2), 102–108.

Kim, S. H., Kim, M. S., Lee, M. S., Park, Y. S., Lee, H. J., Kang, S. A., … Kwon, D. Y. (2016). Korean diet: Characteristics and historical background. *Journal of Ethnic Foods, 3*, 26–31.

Kim, S. J., Dunlap, C. A., Kwon, S. W., & Rooney, A. P. (2015). *Bacillus glycinifermentans* sp. nov., isolated from fermented soybean paste. *International Journal of Systematic and Evolutionary Microbiology, 65*(10), 3586–3590.

Kindoli, S., Lee, H. A., Heo, K., & Kim, J. H. (2012). Properties of a bacteriocin from *Bacillus subtilis* H27 isolated from *Cheonggukjang*. *Food Science and Biotechnology, 21* (6), 1745–1751.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research, 27*(5), 722–736.

Kubo, Y., Rooney, A. P., Tsukakoshi, Y., Nakagawa, R., Hasegawa, H., & Kimura, K. (2011). Phylogenetic analysis of *Bacillus subtilis* strains applicable to natto (fermented soybean) production. *Applied and Environmental Microbiology, 77*(18), 6463–6469.

Kumar, J., Sharma, N., Kaushal, G., Samurailatpam, S., Sahoo, D., Rai, A. K., & Singh, S. P. (2019). Metagenomic insights into the taxonomic and functional features of kinema, a traditional fermented soybean product of Sikkim Himalaya. *Frontiers in Microbiology, 10*, 1744. https://doi.org/10.3389/fmicb.2019.01744

Kwon, D. Y., Chung, K. R., & Jang, D. J. (2019). The history and science of Chongkukjang, a Korean fermented soybean product. *Journal of Ethnic Food, 6*, 5. https://doi.org/10.1186/s42779-019-0004-8

Kwon, Y. S., Lee, S., Lee, S. H., Kim, H. J., & Lee, C. H. (2019). Comparative evaluation of six traditional fermented soybean products in East Asia: A metabolomics approach. *Metabolites, 9*(9), 183. https://doi.org/10.3390/metabo9090183

Lazar, I. M., Karcini, A., Ahuja, S., & Estrada-Palma, C. (2019). Proteogenomic analysis of protein sequence alterations in breast cancer cells. *Scientific Reports, 9*, 10381. https://doi.org/10.1038/s41598-019-46897-z

Lee, C. H., & Kim, M. L. (2016). History of fermented foods in Northeast Asia. In J. P. Tamang (Ed.), *Ethnic Fermented Foods and Alcoholic Beverages of Asia* (pp. 1–16). New Delhi: Springer.

Lee, E. S., Song, Y. J., Kim, K. P., Yim, E. J., Jeong, D. Y., & Cho, S. H. (2015). Manufacturing and quality characteristics of the cheonggukjang fermented using starter derived from rice straw removed *Bacillus cereus* selectively. *Journal of Food Hygiene and Safety, 30*, 65–73.

Lee, H. S. (2013). Diversity of halophilic archaea in fermented foods and human intestines and their application. *Journal of Microbiology and Biotechnology., 23*(12), 1645–1653.

Lee, J. Y., Shim, J. M., Yao, Z., Liu, X., Lee, K. W., Kim, H. J., … Kim, J. H. (2016). Antimicrobial activity of *Bacillus amyloliquefaciens* EMD17 isolated from Cheonggukjang and potential use as a starter for fermented soy foods. *Food Science and Biotechnology, 25*(2), 525–532.

Lee, N. K., Kim, W. S., & Paik, H. D. (2019). *Bacillus* strains as human probiotics: Characterization, safety, microbiome, and probiotic carrier. *Food Science and Biotechnology, 28*(5), 1297–1305.

Lee, Z. S. (1978). Studies on the isolation and characterization of bacteriophage of *Bacillus subtilis* var. 816. *Korean Journal of Microbiology, 16*, 71–78.

Leech, J., Cabrera-Rubio, R., Walsh, A. M., Macori, G., Walsh, C. J., Barton, W., … Cotter, P. D. (2020). Fermented-food metagenomics reveals substrate-associated differences in taxonomy and health-associated and antibiotic resistance determinants. *mSystems, 5*(6), e00522–e620. https://doi.org/10.1128/mSystems.00522-20

Li, X., Zhao, C., Zheng, C., Liu, J., Vu, V. H., Wang, X., & Sun, Q. (2017). Characteristics of microbial community and aroma compounds in traditional fermentation of Pixian broad bean paste as compared to industrial fermentation. *International Journal of Food Properties, 20*(3), S2520–S2531.

Loman, N. J., & Quinlan, A. R. (2014). Poretools: A toolkit for analyzing nanopore sequence data. *Bioinformatics, 30*(23), 3399–3401.

Martino, C., Morton, J. T., Marotz, C. A., Thompson, L. R., Tripathi, A., Knight, R., & Zengler, K. (2019). A novel sparse compositional technique reveals microbial perturbations. *mSystems, 4*(1), e00016–e19. https://doi.org/10.1128/mSystems.00016-19

Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications, 7*(1), 1–9.

Metsalu, T., & Vilo, J. (2015). ClustVis: A web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic Acids Research, 43*(W1), W566–W570.

Moon, J. Y., Kwon, S. W., Hong, S. B., Seok, S. J., Kim, J. S., & Kim, S. J. (2015). Characteristics and functional analysis of *Bacillus* strains from the fermented soybean products, Cheonggukjang. *Korean Journal of Microbiology, 51*(3), 300–307.

Nam, Y. D., Yi, S. H., & Lim, S. I. (2012). Bacterial diversity of cheonggukjang, a traditional Korean fermented food, analysed by barcoded pyrosequencing. *Food Control, 28*(1), 135–142.

Park, D. W., Lee, H. S., Shim, M. S., Yum, K. J., & Seo, J. T. (2020). Do kimchi and cheonggukjang probiotics as a functional food improve androgenetic alopecia? a clinical pilot study. *World Journal of Men's Health, 38*(1), 95–102.

Park, E. J., Kim, K. H., Abell, G. C., Kim, M. S., Roh, S. W., & Bae, J. W. (2011). Metagenomic analysis of the viral communities in fermented foods. *Applied and Environmental Microbiology, 77*(4), 1284–1291.

Park, M., Cho, I., Lee, S., Choi, H. K., Kwon, D. Y., & Kim, Y. S. (2010). Metabolite profiling of Cheonggukjang, a fermented soybean paste, during fermentation by gas chromatography-mass spectrometry and principal component analysis. *Food Chemistry, 122*, 1313–1319.

Park, S., Kim, M. J., Hong, J., Kim, H. J., Yi, S. H., & Lee, M. K. (2014). Selection and characterization of *cheonggukjang* (fast fermented soybean paste)-originated bacterial strains with a high level of s-adenosyl-1-methionine production and probiotics efficacy. *Journal of Medicinal Food, 17*(11), 1170–1176.

Park, Y. K., Jin, Y. H., Lee, J. H., Byun, B. Y., Lee, J., et al. (2020). The role of *Enterococcus faecium* as a key producer and fermentation condition as an influencing factor in tyramine accumulation in *cheonggukjang. Foods, 9*, 915. https://doi.org/10.3390/foods9070915

Parks, D. H., Tyson, G. W., Hugenholtz, P., & Beiko, R. G. (2014). STAMP: Statistical analysis of taxonomic and functional profiles. *Bioinformatics, 30*(21), 3123–3124.

Prakash, T., & Taylor, T. D. (2012). Functional assignment of metagenomic data: Challenges and applications. *Briefings in Bioinformatics, 13*(6), 711–727.

Quince, C., Walker, A., Simpson, J., Loman, N. L., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology, 35*, 833–844.

Ray, K. J., Cotter, S. Y., Arzika, A. M., Kim, J., Boubacar, N., Zhou, Z., … Doan, T. (2019). High-throughput sequencing of pooled samples to determine community-level microbiome diversity. *Annals of Epidemiology, 39*, 63–68.

Rostøl, J. T., & Marraffini, L. (2019). (Ph)ighting phages: How bacteria resist their parasites. *Cell Host and Microbe, 25*(2), 184–194.

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., & Sánchez-Gracia, A. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular Biology and Evolution, 34*(12), 3299–3302.

Samai, P., Pyenson, N., Jiang, W., Goldberg, G. W., Hatoum-Aslan, A., & Marraffini, L. A. (2015). Co-transcriptional DNA and RNA cleavage during type III CRISPR-Cas immunity. *Cell, 161*(5), 1164–1174.

Sapranauskas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., & Siksnys, V. (2011). The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli. Nucleic Acids Research.* https://doi.org/10.1093/nar/gkr606

Scala, G., Serra, A., Marwah, V. S., Saarimäki, L. A., & Greco, D. (2019). FunMappOne: A tool to hierarchically organize and visually navigate functional gene annotations in multiple experiments. *BMC Bioinformatics, 20*, 79. https://doi.org/10.1186/s12859-019-2639-2

Schmieder, R., & Edwards, R. (2011). Quality control and pre-processing of metagenomic datasets. *Bioinformatics, 27*(6), 863–864.

Sevim, V., Lee, J., Egan, R., Clum, A., Hundley, H., Lee, J., … Göker, M. (2019). Shotgun metagenome data of a defined mock community using Oxford Nanopore, PacBio and Illumina technologies. *Scientific Data, 6*(1), 1–9. https://doi.org/10.1038/s41597-019-0287-z

Shapiro, O. H., Kushmaro, A., & Brenner, A. (2010). Bacteriophage predation regulates microbial abundance and diversity in a full-scale bioreactor treating industrial wastewater. *The ISME Journal, 4*(3), 327–336.

Shin, D. H., & Jeong, D. (2015). Korean traditional fermented soybean products: Jang. *Journal of Ethnic Foods, 2*(1), 2–7.

Shin, D. H., Kim, Y. M., Park, W. S., & Kim, J. H. (2016). Ethnic Fermented Foods and Beverages of Korea. In J. P. Tamang (Ed.), *Ethnic Fermented Foods and Alcoholic Beverages of Asia* (pp. 263–308). New Delhi: Springer.

Shin, D. H., Kwon, D. Y., Kim, Y. S., & Jeong, D. Y. (2012). *Science and Technology of Korean Gochujang.* Seoul: Public Health Edu.

Song, Y. S., Jalin, A. M. A. A., Kang, A. R., Lee, C. Y., Park, C. M., Radnabazaar, C., … Kang, S. G. (2010). Thrombolytic activity of Cheonggukjang kinase on the recovery of brain damage in cerebral embolic stroke model. *FASEB Journal, 24*(340), 3. https://doi.org/10.1096/fasebj.24.1_supplement.340.3

Sulaiman, J., Gan, H. M., Yin, W. F., & Chan, K. G. (2014). Microbial succession and the functional potential during the fermentation of Chinese soy sauce brine. *Frontiers in Microbiology., 5*, 556. https://doi.org/10.3389/fmicb.2014.00556

Tamames, J., & Puente-Sánchez, F. (2019). SqueezeMeta, a highly portable, fully automatic metagenomic analysis pipeline. *Frontiers in Microbiology, 9*, 3349. https://doi.org/10.3389/fmicb.2018.03349

Tamang, J. P. (2015). Naturally fermented ethnic soybean foods of India. *Journal of Ethnic Foods, 2*, 8–17.

Tamang, J. P., Cotter, P., Endo, A., Han, N. S., Kort, R., Liu, S. Q., … Hutkins, R. (2020). Fermented foods in a global age: East meets west. *Comprehensive Reviews in Food Science and Food Safety, 19*, 184–217.

Tamang, J. P., Jeyaram, K., Rai, A. K., & Mukherjee, P. K. (2021). Diversity of beneficial microorganisms and their functionalities in community-specific ethnic fermented foods of the Eastern Himalayas. *Food Research International, 148*, 110633. https://doi.org/10.1016/j.foodres.2021.110633

Tamang, J. P., Shin, D. H., Jung, S. J., & Chae, S. W. (2016b). Functional properties of microorganisms in fermented foods. *Frontiers in Microbiology, 7*, 578. https://doi.org/10.3389/fmicb.2016.00578

Tamang, J. P., Watanabe, K., & Holzapfel, W. H. (2016a). Diversity of microorganisms in global fermented foods and beverages. *Frontiers in Microbiology, 7*, 377. https://doi.org/10.3389/fmicb.2016.00377

Weißbecker, C., Buscot, F., & Wubet, T. (2017). Preservation of nucleic acids by freeze-drying for next generation sequencing analyses of soil microbial communities. *Journal of Plant Ecology, 10*(1), 81–90.

Wu, S., Lu, S., Liu, J., Yang, S., Yan, Q., & Jiang, Z. (2020). Physicochemical properties and bioactivities of rice beans fermented by *Bacillus amyloliquefaciens*. *Engineering.* https://doi.org/10.1016/j.eng.2020.10.010

Wu, W. J., & Ahn, B. Y. (2011). Isolation and identification of Bacillus amyloliquefaciens BY01 with high productivity of menaquinone for cheonggukjang production. *Journal of the Korean Society for Applied Biological Chemistry, 54*, 783–789.

Xie, M., Wu, J., An, F., Yue, X., Tao, D., Wu, R., & Lee, Y. (2019). An integrated metagenomic/metaproteomic investigation of microbiota in dajiang-meju, a traditional fermented soybean product in Northeast China. *Food Research International, 115*, 414–424.

Xu, Y. J., Arneja, A. S., Tappia, P. S., & Dhalla, N. S. (2008). The potential health benefits of taurine in cardiovascular disease. *Experimental & Clinical Cardiology, 13*(2), 57–65.

Yang, L., Yang, H. L., Tu, Z. C., & Wang, X. L. (2016). High-throughput sequencing of microbial community diversity and dynamics during douchi fermentation. *PLoS ONE, 11*(12), e0168166. https://doi.org/10.1371/journal.pone.0168166

Ye, Y., & Doak, T. G. (2009). A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Computational Biology, 5*(8), e1000465. https://doi.org/10.1371/journal.pcbi.1000465

Zhang, Z., Schwartz, S., Wagner, L., & Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology, 7*(1–2), 203–214.

Zhao, C. J., Schieber, A., & Gänzle, M. G. (2016). Formation of taste-active amino acids, amino acid derivatives and peptides in food fermentations – A review. *Food Research International, 89*(1), 39–47.

Zhi, Y., Wu, Q., & Xu, Y. (2017). Genome and transcriptome analysis of surfactin biosynthesis in *Bacillus amyloliquefaciens* MT45. *Scientific Reports, 7*, 40976. https://doi.org/10.1038/srep40976