# Conserved accessory genes link phylogenetically distinct *Bacillus subtilis* from Japanese and Indian fermented soyfoods

Kiyohiko Seki[1] & Yukio Nagano[1,2,3,4*]

[1]*Faculty of Agriculture, Saga University, Saga, Japan.*

[2]*The United Graduate School of Agricultural Sciences, Kagoshima University, Kagoshima, Japan*

[3]Vishnu Institute of Pharmaceutical Education & Research, Telangana, India

[4]Shri Vishnu College of Pharmacy, Andhra Pradesh, India

***Corresponding authors:**

E-mail: nagano@cc.saga-u.ac.jp

**Abstract**

*Bacillus subtilis* is central to the production of many fermented soybean foods across Asia, yet the relationship between its phylogeny and niche adaptation remains unclear. To address this, we performed a pangenome analysis on 55 *B. subtilis* strains, including isolates from Japanese natto and Indian bekang. Our analysis revealed a striking evolutionary paradox: a bekang-producing strain, while phylogenetically distant from the homogeneous *B. subtilis* var. *natto* clade, shares a highly conserved accessory gene repertoire with it. This accessory genome profile defines a "natto-type" adaptive strategy, clustering the bekang isolate with the natto strains (Group A) and separating them from other successful soybean-fermenting strains, such as those from Korean cheonggukjang (Group B), which possess a different set of adaptive genes. Analysis of the Group A-specific repertoire revealed an enrichment of transcriptional regulators and metabolic enzymes, representing one successful evolutionary solution for soybean fermentation. The presence of this adaptive module in phylogenetically disparate lineages is consistent with an evolutionary history shaped by events such as horizontal gene transfer. This study highlights a model of polygenic adaptation, in which the acquisition of a large, co-evolved functional module provides a powerful shortcut for rapid adaptation to a complex nutritional niche.

## Introduction

Soybeans are a globally important protein source and have been a primary ingredient in a diverse array of Asian fermented foods for centuries, including staples like soy sauce, miso, tempeh, fǔrǔ, douchi, and natto[1]. Fermentation not only boosts the nutritional value and shelf life of soybeans but also imparts unique flavors and functionalities, profoundly enriching regional food cultures. Central to the production of many of these soybean products is *Bacillus subtilis*[2]. This bacterium plays a pivotal role in creating foods like Japanese natto, bekang from India and Myanmar, pe poke from Myanmar, Korean cheonggukjang, Nepalese kinema, Bhutanese libi ippa, and thua nao from Thailand and Laos. In these foods, *B. subtilis* is crucial for breaking down soy proteins and carbohydrates, thereby generating umami compounds, distinct aromas, and viscous substances such as polyglutamic acid[2].

While *B. subtilis* strains used for Japanese natto (*B. subtilis* var. *natto*) and Korean cheonggukjang have been extensively researched[3,4], our understanding of the genetic diversity, phylogenetic relationships, and adaptation mechanisms of *B. subtilis* from the wide variety of other soybean-fermented foods across Asia remains largely incomplete. The remarkable advancements in genomics now provide access to genome sequence information for many *B. subtilis* strains, particularly those from these diverse Asian origins, enabling detailed inter-strain comparisons[5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23.24,25,26]. To unravel these genetic relationships, pangenome analysis offers a powerful framework[27,28,29]. This method dissects the collective genome of a microbial population into 'core genes' and 'accessory genes'. Core genes, shared by nearly all strains, typically encode essential functions and reflect vertical evolution, making them crucial markers for inferring phylogenetic relationships. In contrast, accessory genes, present only in a subset of strains, often confer traits specific to certain niches—such as environmental adaptation or the production of secondary metabolites—thereby illuminating niche adaptation strategies.

This distinction raises a fundamental question: Do phylogenetic relationships inferred from core genes necessarily align with accessory gene profiles that reflect adaptation to a specific niche, such as the soybean fermentation environment? Strains adapting to similar environments in geographically distant locations might share similar accessory gene sets through mechanisms such as horizontal gene transfer or convergent evolution, even if they belong to distinct core-gene-based phylogenetic lineages. Yet, a comprehensive investigation into this 'phylogeny-versus-functional-gene-pattern' relationship across diverse soybean-fermenting *B. subtilis* strains has been notably lacking.

To address this question, we conducted a pangenome analysis using genomic data from numerous *B. subtilis* strains isolated from various Asian fermented soybean foods. Our primary goal was to evaluate their genetic diversity and identify genetic features linked to their adaptation to the common soybean niche. Through this analysis, we discovered that accessory gene profiles effectively cluster these food-derived strains into distinct groups. Furthermore, we made an unexpected yet central finding: a *B. subtilis* strain from Indian bekang, despite occupying a clearly distinct position on the core-gene-based phylogenetic tree, exhibits a remarkable similarity in its accessory gene repertoire to that of Japanese *B. subtilis* var. *natto* strains.

## Results

### Genomic Data Collection and Pangenome Structure

This study utilized a diverse collection of 55 *B. subtilis* genomes (Table 1). The set includes numerous strains from Asian fermented food production, alongside several environmental and clinical isolates, with the laboratory reference strain 168 included for comparison. Genomes were sourced from public databases or assembled *de novo* from raw sequence reads, and all met high-quality standards for completeness and contamination (Supplementary Table 1). Pangenome analysis of the 55 strains revealed a large and open pangenome, comprising approximately 10,000 genes (Supplementary Figure 1). The pangenome size consistently increased with the addition of new genomes, and the growth curve showed no sign of plateauing, indicating significant genetic diversity within this *B. subtilis* population. Conversely, the core genome stabilized at approximately 3,000 genes, likely representing the essential gene set required for fundamental biological activities within the species.

### Core-Genome Phylogeny Reveals Three Distinct Groups

A phylogenetic tree constructed from the core genome alignment organized the 55 strains into three major groups (Figure 1, Supplementary Figure 2). In our phylogenetic analysis, we designated the clade containing *B. subtilis* var. *natto* strains, used for natto production in Japan, as Group 1. This group formed a tight cluster. It included not only strains from Japanese natto production (Food_JPN_1 to Food_JPN_9) but also strains used for natto production in Korea (Food_KOR_2) and Indonesia (Food_IDN_1). Three other food-isolated strains were also included: Food_KOR_1, Food_CHN_1, and Food_CHN_2. Food_KOR_1 is known to be used in Korean cheonggukjang production, but sufficient information was not available for the latter two. Furthermore, Group 1 contained strains from studies in which *B. subtilis* caused pneumonia (Clin_JPN_1, Clin_JPN_2, and

Clin_JPN_3), strains from microbial collections (Coll_1 to Coll_4), and environmental isolates (Env_CHN_1 to Env_CHN_4, Env_JPN_1). Notably, Env_JPN_1 was isolated from coal-bearing sediment approximately 2.4 km below the ocean floor.

We designated the clade closely related to Group 1 as Group 2. This group comprised *B. subtilis* strains from various Asian fermented soybean foods. While strains used for Korean fermented soybeans like cheonggukjang (Food_KOR_3 to Food_KOR_8, Food_KOR_12) were prominent, Group 2 also included strains from Korean doenjang production (Food_KOR_9, Food_KOR_10) and strains with unknown applications (Food_KOR_11, Food_KOR_13). Furthermore, this group encompassed *B. subtilis* strains from diverse sources outside Korea, including the Indian bekang strain (Food_IND_1), a strain from Myanmar's pe poke (Food_MMR_1), strains from Nepalese and Indian kinema (Food_NPL_1 and Food_IND_2), and a strain from Thai thua nao (Food_THA_1). Within Group 2, a *B. subtilis* strain (Food_NPL_1) used for Nepalese kinema production was most closely related to Group 1.

We designated Group 3 as the cluster positioned outside Groups 1 and 2. With the exceptions of Food_KOR_14 and Food_KOR_16, its members did not form tight clusters among themselves and were all sufficiently distant from Groups 1 and 2 on the phylogenetic tree. Group 3 included the *B. subtilis* reference strain 168 (Coll_5). This group also contained strains used for Korean fermented soybean products such as jeotgal (Food_KOR_14), cheonggukjang (Food_KOR_15), and doenjang (Food_KOR_16), a strain from Laotian thua nao (Food_LAO_1), and isolates from non-soybean fermented foods like Indian chutney (Food_IND_3) and Burkina Faso's maari (Food_BFA_1). Additionally, Group 3 encompassed several environmental isolates (Env_DZA_1, Env_EGY_1, Env_IND_1, Env_USA_1, and Env_USA_2) and a strain from a microbial collection distinct from strain 168 (Coll_6).

**Accessory Gene Profiles Reveal a Functional Grouping that Contradicts Phylogeny**

While the accessory gene repertoires were generally conserved within each phylogenetic group, analysis of the gene presence/absence matrix revealed striking exceptions (Figure 1). Most notably, the *B. subtilis* strain from Indian bekang (Food_IND_1), despite its phylogenetic placement in Group 2, exhibited an accessory gene profile remarkably similar to that of the Group 1 (natto) strains. Based on this discrepancy, we performed a hierarchical clustering analysis using shared gene counts, which redefined the strain groupings based on functional potential rather than phylogeny (Figure 2). This resulted in three new functional groups. The first group, designated Group A, unites the phylogenetic Group 1 strains with the Indian bekang isolate (Food_IND_1). The second, Group B, consists of the strains from phylogenetic Group 2 after excluding both the bekang strain and the Indian beach soil isolate (Env_IND_1). The final cluster, Group C, contains the members of phylogenetic Group 3 along with the Indian beach soil isolate (Env_IND_1). A separate analysis focused only on the 42 strains of Groups A and B further confirmed a clear and distinct separation based on gene presence/absence patterns (Supplementary Figures 3, 4, and 5). This finding underscores that the bekang-producing strain is functionally aligned with *B. subtilis* var. *natto* strains.

**Synteny and Phylogenetic Network Analyses Illuminate Evolutionary History**

To investigate the genomic basis for the shared gene presence/absence patterns, we performed a chromosome-wide synteny analysis comparing representative strains (Figure 3). For this, we required chromosome-level genome sequences and selected the natto strain Food_JPN_7 (Group A), the bekang strain Food_IND_1 (Group A), and the doenjang strain Food_KOR_9 (Group B) as representatives. The results revealed clear synteny among all three strains. However, we found relatively fewer insertions and deletions (indels) between Food_JPN_7 and Food_IND_1, whereas

we detected significantly more indels between the Food_JPN_7/Food_IND_1 pair and Food_KOR_9. This observation held true when comparing with another available chromosome-level Group B strain (Supplementary Figure 6).

Next, we investigated whether complex events like recombination within the core genome could explain the similarity. To explore this, we constructed phylogenetic networks from the core genome alignment for the 42 strains in Groups A and B (Figure 4) and for all 55 strains (Supplementary Figure 7). Using this approach, we did not detect a clear signal of such a relationship between the bekang isolate (Food_IND_1) and the *B. subtilis* var. *natto* strains. However, the analyses did reveal complex, reticulate relationships among several other strains, suggesting that recombination and HGT have occurred within this population. Of particular interest was the evidence of recombination between the bekang isolate and the *B. subtilis* strain (Food_MMR_1) used for pe poke production in Myanmar.

**Identification and Functional Analysis of Group-Specific Accessory Genes**

To characterize the functional differences between the adaptive strategies of Group A and Group B, we extracted their respective specific accessory gene sets. We defined "nearly group-specific genes" using a criterion of ≥80% presence within the target group and ≤20% presence in the comparison group. This approach identified 184 nearly A-specific genes (Supplementary Table 2) and 83 nearly B-specific genes (Supplementary Table 3).

We performed Gene Ontology (GO) and InterPro analyses to investigate the functional characteristics of the Group A-specific gene set. The results of the GO term enrichment analysis for the Biological Process (BP), Molecular Function (MF), and Cellular Component (CC) categories are presented in Figure 5, Figure 6, and Supplementary Figure 8, respectively. A comprehensive table including both GO and InterPro annotations is provided in Supplementary Table 4. Collectively,

these results indicate an enrichment of DNA-binding proteins, enzymes involved in DNA rearrangement, various metabolic enzymes (hydrolases, transferases), and membrane-associated transporters. A curated list of notable genes is presented in Table 2.

Analysis of particularly interesting genes revealed the following: Group A strains typically possess two highly similar copies of *rocB* (arginine degradation)[30], *rssB* (regulator of RpoS)[31,32], and *yveA* (aspartate-proton symporter)[33,34], whereas Group B strains contain only a single copy of each. The *amyE* gene, which encodes an alpha-amylase for starch degradation[35], also showed a duplication pattern in most Group A strains, resulting in two copies (*amyE_1* and *amyE_2*). However, the bekang isolate (Food_IND_1) was a notable exception, containing only a single copy, *amyE_2*. The *bglF* homolog (transport of β-glucosides)[36] in Group A encodes a 609-amino-acid protein, which is significantly longer than the 461-amino-acid version found in Group B. The Group A version of *bglH* (degradation of β-glucosides)[36], at 184 amino acids, appears to be a truncated, likely non-functional protein, in contrast to the full-length, presumably functional version of over 400 amino acids in Group B. Similarly, the *bioF* gene (biotin synthesis) in Group A, at 371 amino acids, is predicted to be non-functional, whereas the Group B homolog, at 389 amino acids, appears to be functional[16,37].

he analysis then focused on the Group B-specific gene set, which comprises genes prevalent in Group B but largely absent from Group A. The results of the GO analysis for the Biological Process (BP), Molecular Function (MF), and Cellular Component (CC) categories for this set are shown in Supplementary Figures 9, 10, and 11, respectively, with detailed annotations provided in Supplementary Table 5. This analysis revealed that Group B possesses its own distinct set of DNA-binding proteins, various metabolic enzymes with hydrolase or transferase activity, and membrane-localized transport enzymes, representing an alternative adaptive toolkit to that of Group A. A selection of characteristic genes from this set is provided in Table 3.

Among the results for the Group B-specific gene set, two genes are of particular interest: *pelB* (Pectate lyase)[38] and *sigV* (sigma factor V)[39]. For *pelB*, no highly similar sequences were detected in Group A. Furthermore, the *pelB* gene exhibited remarkable diversity even within Group B itself. Only the homolog from the *kinema*-producing *B. subtilis* strain (Food_IND_2) appeared to be full-length, encoding a protein of 304 amino acids. In contrast, the versions found in other Group B strains were considerably shorter, ranging from 166 to 170 amino acids, suggesting that they are either truncated or functionally distinct. The other gene of interest, *sigV*, exhibited copy number variation, with Group B strains typically possessing two copies while Group A strains have only one.

The loss of function in the flagellar gene *fliF*, previously noted as a characteristic of *B. subtilis* var. *natto*[16], was observed in 26 natto strains but not in the bekang isolate, indicating it is not a universal Group A trait. Furthermore, the gene presence/absence analysis revealed no differences between Group A and Group B for the key genes associated with polyglutamic acid (PGA) synthesis and its regulation[40]. This set includes the PGA synthesis operon genes (*pgsB*, *pgsC*, and *pgsA*), the PGA-modifying enzyme gene (*pgsE*), and the regulatory genes (*degS*, *degU*, *comP*, *comA*, and *abrB*).

## Discussion

In this study, we demonstrated a striking similarity in the accessory gene repertoire between a *B. subtilis* strain isolated from Indian bekang and the well-characterized *B. subtilis* var. *natto* strains from Japan. This finding is particularly intriguing given that a defining characteristic of Japanese natto is its intense viscosity—a feature produced by polyglutamic acid (PGA) that distinguishes it from most other *B. subtilis*-fermented soybean products. Bekang is a traditional food consumed in Mizoram, India[12,41], and the adjacent Chin State of Myanmar. The shared genomic profile leads to the hypothesis that the bekang-producing lineage may also possess a notable capacity for PGA production. Indeed, a previous study reported substantial PGA production in other bekang-isolated strains, supporting this hypothesis[41]. Although the PGA output was slightly lower than that of elite commercial natto strains, and the strains analyzed were different from the one in our study, the potential for high PGA production presents a compelling candidate for the key functional trait linked to the conserved accessory gene repertoire we have observed.

The most direct explanation for how a bekang-producing strain came to share an adaptive module with distant natto strains is horizontal gene transfer (HGT)[42]. We propose that a "natto-type" adaptive module—a large, co-evolved set of genes—was transferred as a "plug-and-play" system, allowing for rapid adaptation to the challenging nutritional niche of fermenting soybeans[43]. This is highly plausible, as *B. subtilis* is known for its natural competence—the ability to actively take up extracellular DNA[44]—which is abundant in microbially rich environments like soil and fermented foods. While our core genome phylogenetic analysis did not detect a simple recombination event between the lineages, this does not preclude HGT, as such transfers often occur at specific genomic "hotspots" without disrupting the overall core genome phylogeny. Adding another layer of complexity, our network analysis did detect genetic exchange between the bekang strain and a *B. subtilis* strain from pe poke in Myanmar. This suggests the bekang strain may possess a mosaic

genome, combining a core lineage linked to Southeast Asian strains with an adaptive accessory module shared with East Asian natto strains.

An alternative hypothesis involves selective retention from a common ancestor, a process related to convergent evolution[45,46]. This scenario posits that an ancestor of both Group 1 (natto) and Group 2 (bekang) possessed this "soybean-adaptation" gene set, which was subsequently lost in most Group 2 lineages but selectively retained in the bekang and natto lineages due to similar selective pressures.

Our analysis also highlighted the remarkable genetic homogeneity observed among the Japanese *B. subtilis* var. *natto* strains, which form a tight phylogenetic cluster (Group 1). This is primarily attributed to the modern industrial practice of using a limited number of elite starter cultures selected for commercially desirable traits, such as high PGA productivity[3], which created a powerful artificial bottleneck that drastically reduced genetic diversity compared to historical production methods. The situation is made more intriguing by the discovery of phylogenetically similar strains in natural environments, such as the deep-sea isolate Env_JPN_1, indicating that the lineage is not purely an industrial artifact. This can be explained by two non-mutually exclusive hypotheses: the lineage was already naturally homogeneous before the industrial bottleneck was imposed, or highly resilient endospores from industrial strains have dispersed back into diverse ecosystems. Taken together, the extreme genetic similarity of Japanese *B. subtilis* var. *natto* is likely a product of a powerful, recent artificial selection superimposed on a naturally young and relatively homogeneous evolutionary lineage, and the dynamic interplay between these domesticated and wild relatives warrants further investigation.

Functionally, our analysis of the accessory gene sets specific to Group A (natto and bekang) and Group B (cheonggukjang-type) revealed key insights into their adaptive strategies. A prominent feature in the specific accessory gene sets of both groups was the enrichment of helix-turn-helix (HTH) type DNA-binding proteins, suggesting that a sophisticated regulatory toolkit is crucial for

this niche. This included not only standard regulators but also proteins involved in more global gene expression, such as RssB[31,32] and the sigma factor SigV[39]. We also identified numerous metabolic enzymes and transporters, although their specific substrates remain largely unknown. The identification of DNA recombination proteins reflects a history of active genetic exchange, yet the high degree of chromosome-wide synteny observed between the natto and bekang isolates suggests a strong selective pressure to maintain a specific genomic architecture, possibly to ensure the optimal expression of key metabolic pathways. These findings strongly support a model of polygenic adaptation, in which the characteristic traits of these strains are not dictated by a single gene but by a large, conserved set of accessory genes acting in concert. The effect of each individual gene may be modest, but their collective action produces the robust and distinct phenotype.

Finally, this study has a primary limitation: our central finding is based on the analysis of a single bekang isolate. Future studies involving the genomic analysis of a broader collection of *B. subtilis* strains from bekang are crucial to determine whether this conserved accessory repertoire is a common feature. Furthermore, this investigation should be extended to the diverse *B. subtilis* strains from Myanmar's fermented soybean foods. The Bamar people of Myanmar represent another major population—alongside the Japanese and Koreans—with a rich tradition of producing *B. subtilis*-fermented soybean foods, making the region a critical and largely unexplored nexus for this bacterium's diversity. Our discovery of genetic exchange between the Indian bekang and Myanmar pe poke strains further underscores the importance of this region. Such expanded research will be essential to map the complex pathways of gene flow and understand the adaptive radiation of these bacteria across Asia.

## Methods

### Genome Data Acquisition and Quality Control

The 55 *B. subtilis* strains used in this study are detailed in Table 1. For genomes where assemblies were not publicly available, raw reads were retrieved from public read archives and assembled *de novo* using SPAdes (version 4.2.0)[47]. Assembled contigs shorter than 201 bp were subsequently removed using the seq -m 201 option in SeqKit (version 2.8.2)[48]. The quality of all genome sequences was assessed by calculating completeness and contamination values using CheckM2 (version 1.1.0)[49]. To ensure consistent annotation across all datasets, all genome sequences were re-annotated using Prokka (version 1.14.6)[50]. The resulting GFF, GenBank, and FAA files are available from https://drive.google.com/file/d/1dRA0wjLvmyoJHyVJF8F1l4Pl5rchX7GS/view.

### Pangenome and Phylogenetic Analysis

Using the GFF files generated by Prokka as input, a pangenome analysis was conducted with Roary (version 3.13.0)[29] using the options -e -n -v -r. The resulting gene presence/absence matrices for both the full set of 55 samples (Groups A, B, and C; gene_presence_absence_55_samples.csv) and the subset of 42 samples (Groups A and B; gene_presence_absence_42_samples.csv) can be accessed at https://drive.google.com/file/d/1dRA0wjLvmyoJHyVJF8F1l4Pl5rchX7GS/view. Core-gene phylogenetic trees were constructed from the core_gene_alignment_55_samples.aln and core_gene_alignment_42_samples.aln files (available from https://drive.google.com/file/d/1dRA0wjLvmyoJHyVJF8F1l4Pl5rchX7GS/view) generated by Roary. For building a maximum-likelihood phylogenetic tree, IQ-TREE 2 (version 2.2.2.3)[51] was used with the options -m MFP -bb 1000 -alrt 1000. The -m MFP option enables automatic substitution model selection. The resulting phylogenetic tree, including support values, was visualized using FigTree (version 1.4.4) (http://tree.bio.ed.ac.uk/software/figtree/). To generate a

figure that merges the phylogenetic tree with the gene presence/absence matrix, we used a modified version of the roary_plots.py script, which is provided at https://github.com/sanger-pathogens/Roary.git. This modified script, roary_plots_2.py (Supplementary Data), takes the treefile from IQ-TREE 2 and the gene_presence_absence.csv file from Roary as input. The primary modification to this script was the implementation of midpoint rooting for the phylogenetic tree.

**Cluster Analysis Based on Gene Presence/Absence Profiles**

To perform cluster analysis based on accessory gene profiles, the gene_presence_absence_42_samples.csv file generated by Roary was first pre-processed. The initial descriptive annotation columns (columns B through N) were removed. A custom script, count_shared_genes.py (Supplementary Data), was then used to calculate the number of shared genes between all pairs of strains from this matrix, with the results output to a file named shared_counts.csv. Subsequently, this pairwise shared-gene count matrix was used as input for hierarchical clustering. A second custom script, cluster_analysis.py (Supplementary Data), performed the clustering based on Euclidean distances calculated from the shared gene counts. This script also generated a visualization of the results as a combined heatmap and dendrogram, which was then exported in PNG format.

**Visualization of Synteny Blocks**

To visualize synteny blocks among the three selected chromosome-level genomes, Easyfig (version 2.25)[52] was used. This analysis required three GenBank annotation files and two BLAST[53] comparison files as input. The annotation files were the GenBank-formatted outputs generated by Prokka. The two required BLAST files were generated from sequential pairwise blastn searches: the

first search compared genome 1 (query) against genome 2 (database), and the second compared genome 2 (query) against genome 3 (database). For both searches, the output was specified in tabular format using the -outfmt 6 option. Using these three GenBank files and two BLAST files as input, Easyfig was run with its default parameters to generate the comparative visualization.

**Phylogenetic Network Analysis**

A phylogenetic network was constructed from the core gene alignment to investigate potential reticulate evolutionary events, such as recombination. The analysis was performed using SplitsTree (version 6.4.16)[54] with default settings. The input file was the core_gene_alignment_42_samples.aln or core_gene_alignment_55_samples.aln generated by Roary, which was renamed with a .fasta file extension prior to the analysis.

**Identification and Extraction of Group-Specific Accessory Genes**

For the identification of Group A-specific and Group B-specific genes, "nearly group-specific genes" were first defined using a criterion of ≥80% presence within the target group and ≤20% presence in the comparison group. The analysis was performed on the gene_presence_absence_42_samples.csv file from the 42-strain Roary analysis (comprising Groups A and B). This file was first modified by removing the descriptive columns (columns B through N), after which sample names in the header row were prepended with A_ or B_ to denote their respective group affiliations. Next, custom scripts, analyze_genes_A_specific.py and analyze_genes_B_specific.py (Supplementary Data), were executed to parse this modified file and extract the names of all genes that met the defined criteria for Group A and Group B, respectively. The resulting tabular output was then processed with another custom script, process_csv.py (Supplementary Data), to create a list containing only the primary gene names from the leftmost

column. Finally, a third script, process_fasta.py (Supplementary Data), utilized this list of gene names to retrieve the corresponding amino acid sequences in FASTA format from the multiple per-sample protein files (.faa) that were generated by Roary and stored in a specified directory.

**Functional Annotation**

Functional annotation of the Group A-specific and Group B-specific protein sets (.faa files) was performed using OmicsBox (version 3.4.6; GO version March 16, 2025). The workflow consisted of several integrated steps. First, a similarity search was conducted against the Diamond BLAST database (2024-11-20 release) using the blastp mode[55]. This search was taxonomically filtered to *Bacillus* (taxid: 1386) with an E-value cutoff of 1.0E-3. Next, protein domains and families were identified by running InterProScan (version 5.72-103.0)[56]. Finally, GO terms were mapped and assigned to each sequence within OmicsBox based on the combined results of the previous steps[57,58,59]. For the final annotation, results were filtered with an E-value-Hit Filter of 1.0E-6 and a Hit Filter of 500.

**Use of Generative AI**

The generative AI models Gemini and ChatGPT were used to improve the language and clarity of the manuscript and to assist in writing the Python scripts.

## References


1. Harahap, I. A., Suliburska, J., Karaca, A. C., Capanoglu, E. & Esatbeyoglu, T. Fermented soy products: a review of bioactives for health from fermentation to functionality. *Compr. Rev. Food Sci. Food Saf.* **24**, e70080; 10.1111/1541-4337.70080 (2025).
2. Gopikrishna, T., Suresh Kumar, H. K., Perumal, K. & Elangovan, E. Impact of *Bacillus* in fermented soybean foods on human health. *Ann. Microbiol.* **71**, 1–16 (2021).
3. Miyazawa, T. et al. Biological functions of compounds from *Bacillus subtilis* and its subspecies, *Bacillus subtilis* natto. *Food Bioengineering* **1**, 241–251 (2022).
4. Kim, I. S., Hwang, C. W., Yang, W. S. & Kim, C. H. Current perspectives on the physiological activities of fermented soybean-derived cheonggukjang. *Int. J. Mol. Sci.* **22**, 5746; 10.3390/ijms22115746 (2021).
5. Tani, T. et al. A case of community-acquired pneumonia caused by *Bacillus subtilis* subsp. *natto* in an immunocompetent patient. *Respirol. Case Rep.* **12**, e01384; 10.1002/rcr2.1384 (2024).
6. Tan, S., Meng., Y, Su., A, Zhang., C & Ren, Y. Draft genome sequence of *Bacillus subtilis* subsp. *natto* strain CGMCC 2108, a high producer of poly-γ-glutamic acid. *Genome Announc.* **4**, e00510-16; 10.1128/genomea.00426-16 (2016).
7. Wang, F., Gong, L., Zhou, L. & Liang, J. Complete genome sequence of the poly-γ-glutamate-synthesizing bacterium *Bacillus subtilis* Bs-115. *Genome Announc.* **6**, e00318-18; 10.1128/genomea.00197-18 (2018).
8. Dong, H., Chang, J., He, X., Hou, Q. & Long, W. Complete genome sequence of *Bacillus subtilis* strain CGMCC 12426, an efficient poly-γ-glutamate producer. *Genome Announc.* **5**, e01168-17; 10.1128/genomea.01163-17 (2017).

9. Wei, Y., Cao, J., Fang, J., Kato, C. & Cui, W. Complete genome sequence of *Bacillus subtilis* strain 29R7-12, a piezophilic bacterium isolated from coal-bearing sediment 2.4 kilometers below the seafloor. *Genome Announc.* **5**, e01728-16; 10.1128/genomea.01621-16 (2017).
10. Min, X. et al. Comparative transcriptome analysis reveals the underlying mechanism for over-accumulation of menaquinone-7 in *Bacillus subtilis* natto mutant. *Biochem. Eng. J.* **174**, 108097; 10.1016/j.bej.2021.108097 (2021).
11. Dikson, D. *et al.* Whole-genome analysis of *Bacillus subtilis* G8 isolated from natto. *Biodivers. J. Biol. Divers.* **23,** 1378–1385 (2022).
12. Kumar, H. K. S., Gopikrishna, T., Tapasya, K., Perumal, K. & Elavarashi, E. A computational genome analysis of strain *Bacillus subtilis* MIZ-8 isolated from bekang reveals a distinct chromosome and plasmid conferring selective advantage. *J. Microbiol. Biotechnol. Food Sci.* **13,** e9141; 10.55251/jmbfs.9141 (2023).
13. Zhu, J. *et al.* Genomic characterization and related functional genes of γ-polyglutamic acid producing *Bacillus subtilis*. *BMC Microbiol.* **24,** 125; 10.1186/s12866-024-03262-z (2024).
14. Kamada, M. *et al.* Whole genome complete resequencing of *Bacillus subtilis* natto by combining long reads with high-quality short reads. *PLoS ONE* **9,** e109999; 10.1371/journal.pone.0109999 (2014).
15. Kubo, Y. *et al.* Complete genome sequence of *Bacillus subtilis* subsp. *natto* NARUSE using PacBio sequencing. *Microbiol. Resour. Announc.* **12,** e00578-23; 10.1128/MRA.00578-23 (2023).
16. Kamada, M. *et al.* Whole-genome sequencing and comparative genome analysis of *Bacillus subtilis* strains isolated from non-salted fermented soybean foods. *PLoS ONE* **10,** e0141369; 10.1371/journal.pone.0141369 (2015).

17. Parks, D. *et al.* Genome sequence of *Bacillus subtilis* natto VK161, a novel strain that produces vitamin K2. *Microbiol. Resour. Announc.* **8,** e00854-19; 10.1128/mra.00444-19 (2019).
18. Tamang, J. P., Kharnaior, P. & Pariyar, P. Whole genome sequencing of the poly-γ-glutamic acid-producing novel *Bacillus subtilis* Tamang strain, isolated from spontaneously fermented kinema. *Food Res. Int.* **190,** 114655; 10.1016/j.foodres.2024.114655 (2024).
19. Bang, M. S. *et al.* Complete genome sequence of *Bacillus subtilis* strain DKU_NT_02, isolated from traditional Korean food using soybean (chung-gook-jang) for high-quality poly-γ-glutamic acid activity. *Genome Announc.* **6,** e00525-18; 10.1128/genomea.00525-18 (2018).
20. Jeong, H. W. *et al.* Complete genome sequence of *Bacillus subtilis* strain DKU_NT_03, isolated from a traditional Korean food using soybean (Chung-gook-jang) for high-quality nattokinase activity. *Genome Announc*. **6,** e00526-18; 10.1128/genomea.00526-18 (2018).
21. Choi, J. W. *et al.* Engineering *Bacillus subtilis* J46 for efficient utilization of galactose through adaptive laboratory evolution. *AMB Express* **14,** 14; 10.1128/genomea.00526-18 (2024).
22. Kunst, F. J. *et al.* The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390,** 249–256 (1997).
23. Choubane, S., Gabed, N., Khelil, O. & Cheba, B. A. Draft genome sequence of *Bacillus subtilis* TLO3, isolated from olive tree rhizosphere soil. *Microbiol. Resour. Announc.* **7,** e01128-18; 10.1128/mra.00852-18 (2018).
24. Hamdy, A. A. *et al.* Complete genome sequence and comparative analysis of two potential probiotics *Bacillus subtilis* isolated from honey and honeybee microbiomes. *J. Genet. Eng. Biotechnol.* **18,** 34; 10.1186/s43141-020-00050-w (2020).
25. Pramod, S. *et al.* Data on the genome of *Bacillus subtilis* A1-Midalam from beach soil. *Data Brief* **39,** 107552; 10.1016/j.dib.2021.107552 (2021).

26. Epparti, P., Eligar, S. M., Sattur, A. P., Kumar, B. G. & Halami, P. M. Characterization of dual bacteriocins producing *Bacillus subtilis* SC3.7 isolated from fermented food. *LWT* **154,** 112854; 10.1016/j.lwt.2021.112854 (2022).
27. Golicz, A. G., Bayer, P. E., Bhalla, P. L., Batley, J. & Edwards, D. Computational pan-genomics: status, promises and challenges. *Brief. Bioinform.* **21,** 1–19 (2020).
28. Verma, D., Satyanarayana, T. & Dias, P. J. Microbial comparative genomics and pangenomics: new tools, approaches and insights into gene and genome evolution. *Front. Genet.* **15,** 1490645; 10.3389/fgene.2024.1490645 (2024).
29. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31,** 3691–3693 (2015).
30. Warneke, R. *et al.* Ornithine is the central intermediate in the arginine degradative pathway and its regulation in *Bacillus subtilis*. *J. Biol. Chem.* **299,** 104944; 10.1016/j.jbc.2023.104944 (2023).
31. Bouillet, S., Hamdallah, I., Majdalani, N., Tripathi, A. & Gottesman, S. A negative feedback loop is critical for recovery of RpoS after stress in *Escherichia coli*. *PLoS Genet.* **20,** e1011059; 10.1371/journal.pgen.1011059 (2024).
32. Battesti, A. *et al.* Anti-adaptors provide multiple modes for regulation of the RssB adaptor protein. *Genes Dev.* **27,** 2722–2735 (2013).
33. Lorca, G., Winnen, B. & Saier, M. H. Jr. Identification of the L-aspartate transporter in *Bacillus subtilis*. *J. Bacteriol.* **185,** 3218–3222 (2003).
34. Zaprasis, A., Bleisteiner, M., Kerres, A., Hoffmann, T. & Bremer, E. Uptake of amino acids and their metabolic conversion into the compatible solute proline confers osmoprotection to *Bacillus subtilis*. *Appl. Environ. Microbiol.* **81,** 250–259 (2015).

35. Murphy, N., McConnell, D. J. & Cantwell, B. A. The DNA sequence of the gene and genetic control sites for the excreted *B. subtilis* enzyme β-glucanase. *Nucleic Acids Res.* **12,** 5355–5367 (1984).

36. Le Coq, D., Lindner, C., Krüger, S., Steinmetz, M. & Stülke, J. New β-glucoside (*bgl*) genes in *Bacillus subtilis*: the *bglP* gene product has both transport and regulatory functions similar to those of BglF, its *Escherichia coli* homolog. *J. Bacteriol.* **177,** 1527–1535 (1995).

37. Sasaki, M., Kawamura, F. & Kurusu, Y. Genetic analysis of an incomplete *bio* operon in a biotin auxotrophic strain of *Bacillus subtilis* natto OK2. *Biosci. Biotechnol. Biochem.* **68,** 739–742 (2004).

38. Nasser, W., Awade, A. C., Reverchon, S. & Robert-Baudouy, J. Pectate lyase from *Bacillus subtilis*: molecular characterization of the gene, and properties of the cloned enzyme. *FEBS Lett.* **335,** 319–326 (1993).

39. Ho, T. D., Hastie, J. L., Intile, P. J. & Ellermeier, C. D. The *Bacillus subtilis* extracytoplasmic function σ factor σV is induced by lysozyme and provides resistance to lysozyme. *J. Bacteriol.* **193,** 6215–6222 (2011).

40. Elbanna, K., Alsulami, F. S., Neyaz, L. A. & Abulreesh, H. H. Poly-γ-glutamic acid: a unique microbial biopolymer with diverse commercial applicability. *Front. Microbiol.* **15,** 1348411; 10.3389/fmicb.2024.1348411 (2024).

41. Chettri, R. & Tamang, J. P. Functional properties of Tungrymbai and Bekang, naturally fermented soybean foods of North East India. *J. Ethn. Foods* **1,** 8–14 (2014).

42. Arnold, B. J., Huang, I. T. & Hanage, W. P. Horizontal gene transfer and adaptive evolution in bacteria. *Nat. Rev. Microbiol.* **20,** 206–218 (2022).

43. Good, B. H., Bhatt, A. S. & McDonald, M. J. Unraveling the tempo and mode of horizontal gene transfer in bacteria. *Trends Microbiol.* **32,** 52–64 (2024).

44. Spizizen, J. Transformation of biochemically deficient strains of *Bacillus subtilis* by deoxyribonucleate. *Proc. Natl. Acad. Sci. U. S. A.* **44,** 1072–1078 (1958).

45. Albalat, R. & Cañestro, C. Evolution by gene loss. *Nat. Rev. Genet.* **17,** 379–391 (2016).

46. Mira, A., Ochman, H. & Moran, N. A. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17,** 589–596 (2001).

47. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes de novo assembler. *Curr. Protoc. Bioinformatics* **70,** e102; 10.1002/cpbi.102 (2020).

48. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE* **11,** e0163962; 10.1371/journal.pone.0163962 (2016).

49. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20,** 1203–1212 (2023).

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

51. Minh, B. Q. *et al.* IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37,** 1530–1534 (2020).

52. Sullivan, M. J., Petty, N. K. & Beatson, S. A. Easyfig: a genome comparison visualizer. *Bioinformatics* **27,** 1009–1010 (2011).

53. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).

54. Huson, D. H. & Bryant, D. The SplitsTree app: interactive analysis and visualization using phylogenetic trees and networks. *Nat. Methods* **21**, 1773–1774 (2024).

55. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18,** 366–368 (2021).

56. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30,** 1236–1240 (2014).

57. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nat. Genet.* **25,** 25–29 (2000).

58. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51,** D523–D531 (2023).

59. Götz, S. *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36,** 3420–3435 (2008).

**Acknowledgements**

This study was supported in part by the Institute for Fermentation, Osaka, Japan (IFO). We are especially grateful to Dr. Myat Htoo San of the National Institute of Advanced Industrial Science and Technology (AIST), Hokkaido, Japan, and her friend, Ms. Zing Cin Tial, a member of the Chin community from Hakha, Chin State, Myanmar. They kindly provided valuable information and cultural insights into the traditional production of bekang in Chin State, Myanmar, for which there is little published literature. We also thank Drs. Taiki Futagami and Hisanori Tamaki of the Faculty of Agriculture, Kagoshima University, and Drs. Masatoshi Goto and Genta Kobayashi of the Faculty of Agriculture, Saga University, for guiding us into the research of fermented foods.

**Author Contributions**

K.S. and Y.N. conceived and designed the study, performed the bioinformatic analyses, and interpreted the data. Y.N. wrote the original manuscript draft, and K.S. reviewed and edited the manuscript. Both authors read and approved the final manuscript.

**Data Availability Statement**

The primary genome sequence data analyzed in this study were retrieved from publicly available databases; specific accession numbers for all strains are detailed in Table 1. The key data files generated during the computational analysis, including the gene presence/absence matrices, can be accessed at https://drive.google.com/file/d/1dRA0wjLvmyoJHyVJF8F1l4Pl5rchX7GS/view. All custom Python scripts developed for this work are available as Supplementary Data.

**Competing Interests**

Kiyohiko Seki and Yukio Nagano conduct collaborative research with Marumiya, Inc., a food manufacturer that produces natto. Yukio Nagano also conducts collaborative research with Senoo Suisan, Inc., and Sante Laboratories, Inc. The authors declare that these collaborations are entirely separate from the present study. Specifically, the collaborative research with Marumiya concerns different aspects of natto production and had no role in the study design, data collection and analysis, decision to publish, or preparation of this manuscript.

## Figure Legends

**Figure 1.** Core-genome-based phylogeny and accessory gene presence/absence matrix of the 55 *B. subtilis* strains. The phylogenetic tree on the left was constructed from a core genome alignment of 55 strains using the maximum likelihood method and is presented as midpoint-rooted. The matrix on the right, generated by Roary, visualizes the pangenome, which consists of 9,848 gene clusters. Each row corresponds to a strain in the tree, and each column represents a single gene cluster. Dark blue cells indicate the presence of a gene cluster in a strain, while white cells indicate its absence. Group assignments for each strain are denoted between the tree and the matrix using colored bars and alphanumeric labels, corresponding to the phylogenetic groups (1, 2, 3) and the accessory-genome-based functional groups (A, B, C) as defined in the text.

**Figure 2.** Hierarchical clustering of the 55 *B. subtilis* strains based on shared gene profiles. The heatmap shows the pairwise number of shared genes between all strains, with the color intensity corresponding to the values on the scale bar. Hierarchical clustering was performed using Euclidean distance based on these shared gene counts. The labels next to the dendrograms indicate the functional group (A, B, or C) and phylogenetic group (1, 2, or 3) for each strain.

**Figure 3.** Chromosome-wide synteny analysis of representative strains from Group A and Group B. This visualization compares the complete chromosome sequences of three strains, represented by horizontal bars: *B. subtilis* var. *natto* BEST195 (Food_JPN_7, top bar), the bekang isolate (Food_IND_1, middle bar), and *B. subtilis* HRBS-10TDI13 (Food_KOR_9, bottom bar). The shaded regions connecting the genomes denote areas of nucleotide similarity identified by BLASTn, with the grayscale intensity corresponding to the percentage identity shown in the scale bar. The top panel shows the pairwise comparison between the natto strain (Food_JPN_7) and the bekang isolate

(Food_IND_1). The bottom panel shows the pairwise comparison between the bekang isolate (Food_IND_1) and the Group B strain (Food_KOR_9).

**Figure 4.** Phylogenetic network of 42 *B. subtilis* strains from Groups A and B. The network was constructed from the core genome alignment using the Neighbor-Net algorithm. The tree-like portions of the network depict vertical, clonal inheritance, while the reticulate (box-like) structures indicate conflicting phylogenetic signals suggestive of recombination or horizontal gene transfer events.

**Figure 5.** Gene Ontology (GO) enrichment analysis for the Biological Process (BP) category of the Group A-specific gene set. The figure shows a Directed Acyclic Graph of enriched GO terms, generated using OmicsBox. Each node (box) represents a GO term and displays its ID, description, NodeScore, and the number of associated gene sequences. The color intensity of each node is proportional to its enrichment significance, with more significantly enriched terms shown in shades of orange. The connecting lines represent hierarchical relationships between the terms.

**Figure 6.** Gene Ontology (GO) enrichment analysis for the Molecular Function (MF) category of the Group A-specific gene set. The figure shows a Directed Acyclic Graph of enriched GO terms. Each box represents a GO term, displaying its ID, description, and the number of associated gene sequences. The color intensity of the boxes is proportional to the significance of enrichment, with more significant terms shown in shades of orange. The connecting lines represent hierarchical relationships between the terms.

585 **Table 1.** Details of the publicly available data used in this study.

| Name | Functional Group | Phylogenetic Group | Accession number | Strain name | Origin details | Location | Publication |
|---|---|---|---|---|---|---|---|
| Clin_JPN_1 | A | 1 | GCF_037094835.1 | TUM22925 | Patient blood | Japan | 5 |
| Clin_JPN_2 | A | 1 | GCF_037094855.1 | TUM22926 | Patient bronchoalveolar lavage fluid | Japan | 5 |
| Clin_JPN_3 | A | 1 | GCF_037095795.1 | TUM22927 | The natto with the patient usually eating | Japan | 5 |
| Coll_1 | A | 1 | GCF_002982175.1 | ATCC 21228 | N/A | N/A | N/A |
| Coll_2 | A | 1 | GCF_024508195.1 | BGSC 27E1 | N/A | N/A | N/A |
| Coll_3 | A | 1 | GCF_001565875.1 | CGMCC 2108 | Soil | N/A | 6 |
| Coll_4 | A | 1 | GCF_024507815.1 | BGSC 27E3 | N/A | N/A | N/A |
| Env_CHN_1 | A | 1 | GCF_002096095.1 | Bs-115 | Soil of a corn field | Shandong, China | 7 |
| Env_CHN_2 | A | 1 | GCF_045343855.1 | G01 | N/A | Guangdong, China | N/A |
| Env_CHN_3 | A | 1 | GCF_001890405.1 | KH2 | Soil | Tianjin, China | 8 |
| Env_CHN_4 | A | 1 | GCF_023278545.1 | ZHX3 | Activated sludge | Changsha, China | N/A |
| Env_JPN_1 | A | 1 | GCF_001902555.1 | 29R7-12 | Coal Bed | Shimokita Peninsula, Japan | 9 |
| Food_CHN_1 | A | 1 | GCF_022701095.1 | BN0 | Soybean product | Wuxi, China | 10 |
| Food_CHN_2 | A | 1 | GCF_020773035.1 | SCP010-1 | Bean Paste | Sichuan, China | N/A |
| Food_IDN_1 | A | 1 | GCF_027923485.1 | G8 | Natto | Indonesia | 11 |
| Food_IND_1 | A | 2 | GCF_018986855.1 | Miz-8 | Bekang | Mizoram, India | 12 |
| Food_JPN_1 | A | 1 | GCF_003665335.1 | N1-1 | Natto | Japan | N/A |
| Food_JPN_2 | A | 1 | GCF_003665315.1 | N2-2 | Natto | Japan | N/A |
| Food_JPN_3 | A | 1 | GCF_003665355.1 | N3-1 | Natto | Japan | N/A |
| Food_JPN_4 | A | 1 | GCF_033412715.1 | N3378-2at | Natto | Hokkaido, Japan | 13 |
| Food_JPN_5 | A | 1 | GCF_033413615.1 | N3378-3At | Natto | Hokkaido, Japan | 13 |
| Food_JPN_6 | A | 1 | GCF_003665295.1 | N4-2 | Natto | Japan | N/A |
| Food_JPN_7 | A | 1 | GCF_000209795.2 | BEST195 | Natto | Japan | 14 |
| Food_JPN_8 | A | 1 | GCF_030297075.1 | NARUSE | Natto | Japan | 15 |
| Food_JPN_9 | A | 1 | Assembled from DRR029699* | TAKAHASHI | Natto | Japan | 16 |
| Food_KOR_1 | A | 1 | GCF_014961985.1 | CMIN-4 | Cheonggukjang | Jeonbuk, South Korea | N/A |
| Food_KOR_2 | A | 1 | GCF_004323065.1 | VK161 | Natto | Seongnam, South Korea | 17 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Food_IND_2 | B | 2 | GCF_037169145.1_ | Tamang | Kinema | Sikkim, India | 18 |
| Food_KOR_3 | B | 2 | GCF_003665395.1 | 2KL1 | Cheonggukjang | Jeonju, South Korea | N/A |
| Food_KOR_4 | B | 2 | GCF_003665275.1 | 2RL2-3 | Cheonggukjang | Gyeonggi, South Korea | N/A |
| Food_KOR_5 | B | 2 | Assembled from DRR029692* | NFRI8338 | Cheonggukjang | Seoul, South Korea | 16 |
| Food_KOR_6 | B | 2 | GCF_002269175.1 | DKU_NT_02 | Cheonggukjang | Cheonan, South Korea | 19 |
| Food_KOR_7 | B | 2 | GCF_002269195.1 | DKU_NT_03 | Cheonggukjang | Cheonan, South Korea | 20 |
| Food_KOR_8 | B | 2 | GCF_003665195.1 | GFR-12 | Cheonggukjang | Gyeonggi, South Korea | N/A |
| Food_KOR_9 | B | 2 | GCF_001747445.1 | HRBS-10TDI13 | Doenjang | Sunchang, South Korea | N/A |
| Food_KOR_10 | B | 2 | GCF_033113625.1 | J46 | Doenjang | South Korea | 21 |
| Food_KOR_11 | B | 2 | GCF_046840395.1 | KFRI-P74 | Fermented soybean paste | Ansung, South Korea | N/A |
| Food_KOR_12 | B | 2 | GCF_003665235.1 | MH-1 | Cheonggukjang | Chungcheongbuk-do, South Korea | N/A |
| Food_KOR_13 | B | 2 | GCF_035621695.1_ | PN1236 | Fermented soybean food | South Korea | N/A |
| Food_MMR_1 | B | 2 | Assembled from DRR029694* | NFRI8316 | Pe poke | Yangon, Myanmar | 16 |
| Food_NPL_1 | B | 2 | Assembled from DRR029696* | NFRI8292 | Kinema | Hille, Nepal | 16 |
| Food_THA_1 | B | 2 | Assembled from DRR029695* | NFRI8347 | Thua nao | Chiang Mai, Thailand | 16 |
| Coll_5 | C | 3 | GCF_000009045.1 | 168 | N/A | N/A | 22 |
| Coll_6 | C | 3 | GCF_004101945.1 | ATCC 11774 | N/A | N/A | N/A |
| Env_DZA_1 | C | 3 | GCF_002142595.1 | TLO3 | Olive tree rhizosphere | Tlemcen, Algeria | 23 |
| Env_EGY_1 | C | 3 | GCF_009497795.1 | MENO2 | Honeybee gut | Egypt | 24 |
| Env_IND_1 | C | 2 | GCF_018454405.1 | A1 - Midalam | Beach soil | Tamil Nadu, India | 25 |
| Env_USA_1 | C | 3 | GCF_037976695.1 | FSL K6-1366 | HTST Pasteurization | N/A | N/A |
| Env_USA_2 | C | 3 | GCF_047302535.1 | SY1483 | Root | USA | N/A |
| Food_BFA_1 | C | 3 | GCA_029536855.2 | PRO120 | Maari | Toulfe, Burkina Faso | N/A |
| Food_IND_3 | C | 3 | GCF_018224865.2 | SC3.7 | Soybean seed Chutney | Manipur, India | 26 |
| Food_KOR_14 | C | 3 | GCF_030285965.1 | KFRI-FA111 | Jeotgal | South Korea | N/A |
| Food_KOR_15 | C | 3 | GCF_003665215.1 | PJ-7 | Cheonggukjang | Gyeonggi, South Korea | N/A |
| Food_KOR_16 | C | 3 | GCF_009913275.1 | SRCM102754 | Doenjang | South Korea | N/A |
| Food_LAO_1 | C | 3 | Assembled from DRR029693* | NFRI8302 | Thua nao | Ruanpanpan, Laos | 16 |

*Assembled *de novo* from Sequence Read Archive data. Abbreviations used in "Name": Clin – Clinical study; Coll – Collection; Env – Environmental; JPN – Japan; CHN – China; IDN – Indonesia; IND – India; KOR – South Korea; MMR – Myanmar; NPL – Nepal; THA – Thailand; DZA – Algeria; EGY – Egypt; BFA – Burkina Faso; LAO – Laos.

588 **Table 2.** A selection of nearly A-specific genes.

| SeqName[1] | The number in Group A[2] | Presence in Bekang strain (1/0) | The number in Group B | Description[3] | GO IDs | GO Names | Enzyme Codes | Enzyme Names |
|---|---|---|---|---|---|---|---|---|
| *xerC_6* | 27 | 1 | 2 | site-specific integrase | P:GO:0006310; F:GO:0003677 | P:DNA recombination; F:DNA binding | | |
| *xerC_9* | 27 | 1 | 1 | tyrosine-type recombinase/integrase | P:GO:0006310; F:GO:0003677 | P:DNA recombination; F:DNA binding | | |
| *xerD_2* | 27 | 1 | 2 | tyrosine-type recombinase/integrase | P:GO:0006310; F:GO:0003677 | P:DNA recombination; F:DNA binding | | |
| *xerD_3* | 27 | 1 | 1 | site-specific integrase | P:GO:0006310; F:GO:0003677 | P:DNA recombination; F:DNA binding | | |
| group_2454 | 27 | 1 | 0 | site-specific integrase | P:GO:0006310; F:GO:0003677 | P:DNA recombination; F:DNA binding | | |
| group_4906 | 27 | 1 | 3 | transposase | P:GO:0006310; F:GO:0003677; F:GO:0140097 | P:DNA recombination; F:DNA binding; F:catalytic activity, acting on DNA | | |
| group_6181 | 22 | 1 | 1 | IS21 family transposase | P:GO:0006310; F:GO:0003677; F:GO:0140097 | P:DNA recombination; F:DNA binding; F:catalytic activity, acting on DNA | | |
| group_1396 | 25 | 0 | 3 | helix-turn-helix transcriptional regulator | F:GO:0003677 | F:DNA binding | | |
| group_1783 | 27 | 1 | 0 | ERCC4 domain-containing protein | F:GO:0003677; F:GO:0003824 | F:DNA binding; F:catalytic activity | | |
| group_1787 | 27 | 1 | 1 | host-nuclease inhibitor Gam family protein | F:GO:0003677 | F:DNA binding | | |
| group_1890 | 27 | 1 | 1 | helix-turn-helix transcriptional regulator | F:GO:0003677 | F:DNA binding | | |
| group_2132 | 27 | 1 | 2 | helix-turn-helix domain-containing protein | F:GO:0003677 | F:DNA binding | | |
| group_2184 | 27 | 1 | 2 | helix-turn-helix transcriptional | F:GO:0003677 | F:DNA binding | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | regulator | | | | |
| group_2238 | 27 | 1 | 1 | helix-turn-helix transcriptional regulator | F:GO:0003677 | F:DNA binding | | |
| group_2240 | 27 | 1 | 1 | FtsK/SpoIIIE domain-containing protein | F:GO:0003677 | F:DNA binding | | |
| group_2450 | 27 | 1 | 0 | phage regulatory protein/antirepressor Ant | F:GO:0003677 | F:DNA binding | | |
| group_2451 | 27 | 1 | 0 | helix-turn-helix transcriptional regulator | F:GO:0003677 | F:DNA binding | | |
| group_2452 | 27 | 1 | 0 | helix-turn-helix transcriptional regulator | F:GO:0003677 | F:DNA binding | | |
| group_2456 | 27 | 1 | 0 | helix-turn-helix transcriptional regulator | F:GO:0003677 | F:DNA binding | | |
| group_2457 | 27 | 1 | 0 | helix-turn-helix domain-containing protein | F:GO:0003677 | F:DNA binding | | |
| group_2531 | 27 | 1 | 0 | helix-turn-helix transcriptional regulator | F:GO:0003677 | F:DNA binding | | |
| group_2546 | 27 | 1 | 3 | helix-turn-helix transcriptional regulator | F:GO:0003677 | F:DNA binding | | |
| group_2211 | 27 | 1 | 3 | MarR family transcriptional regulator | P:GO:0006355; F:GO:0140110 | P:regulation of DNA-templated transcription; F:transcription regulator activity | | |
| group_2518 | 27 | 1 | 1 | sigma factor-like helix-turn-helix DNA-binding protein | P:GO:0006355; F:GO:0140110 | P:regulation of DNA-templated transcription; F:transcription regulator activity | | |
| group_5156 | 26 | 1 | 1 | IS3 family transposase | P:GO:0015074; F:GO:0003676 | P:DNA integration; F:nucleic acid binding | | |
| *insK* | 23 | 1 | 1 | IS3 family transposase | P:GO:0015074; F:GO:0003676 | P:DNA integration; F:nucleic acid binding | | |
| *amyE_1* | 25 | 0 | 3 | Alpha-amylase | F:GO:0016787 P:GO:0005975 | F:hydrolase activity P:carbohydrate metabolic | EC:3.2.1.1 | alpha-amylase |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | process | | |
| *bglH_2* | 26 | 1 | 3 | Aryl-phospho-beta-D-glucosidase BglH | F:GO:0016787<br>P:GO:0005975 | F:hydrolase activity<br>P:carbohydrate metabolic process | EC:3.2.1 | Glycosylases |
| *cheC_1* | 26 | 1 | 0 | CheY-P phosphatase CheC | F:GO:0016787 | F:hydrolase activity | EC:3 | Hydrolases |
| *cwlA_2* | 27 | 1 | 1 | N-acetylmuramoyl-L-alanine amidase | F:GO:0016787<br>P:GO:1901135; | F:hydrolase activity<br>P:carbohydrate derivative metabolic process; | EC:3.5.1.28 | N-acetylmuramoyl-L-alanine amidase |
| *lagD* | 27 | 1 | 3 | peptide cleavage/export ABC transporter | P:GO:0055085;<br>F:GO:0005215;<br>F:GO:0016787;<br>F:GO:0140096;<br>F:GO:0140657 | P:transmembrane transport;<br>F:transporter activity;<br>F:hydrolase activity;<br>F:catalytic activity, acting on a protein;<br>F:ATP-dependent activity | EC:3.4;<br>EC:7.6.2.2;<br>EC:7.2.2;<br>EC:3.6.1.15 | Acting on peptide bonds (peptidases);<br>ABC-type xenobiotic transporter;<br>Catalysing the translocation of inorganic cations;<br>nucleoside-triphosphate phosphatase |
| *rocB_1* | 25 | 1 | 2 | M20/M25/M40 family metallo-hydrolase | F:GO:0016787 | F:hydrolase activity | EC:3 | Hydrolases |
| *uvrD* | 22 | 1 | 3 | ATP-dependent helicase | F:GO:0016787 | F:hydrolase activity | EC:3 | Hydrolases |
| *yesZ_1* | 26 | 0 | 1 | Beta-galactosidase YesZ | F:GO:0016787<br>P:GO:0005975 | F:hydrolase activity<br>P:carbohydrate metabolic process | EC:3.2.1.23 | beta-galactosidase |
| *yesZ_2* | 27 | 1 | 1 | Beta-galactosidase YesZ | F:GO:0016787<br>P:GO:0005975 | F:hydrolase activity<br>P:carbohydrate metabolic process | EC:3.2.1.23 | beta-galactosidase |
| group_689 | 27 | 1 | 3 | ABC transporter ATP-binding protein | P:GO:0055085;<br>F:GO:0005215;<br>F:GO:0016787;<br>F:GO:0140657 | P:transmembrane transport;<br>F:transporter activity;<br>F:hydrolase activity;<br>F:ATP-dependent activity | EC:7.2.2;<br>EC:3.6.1.15 | Catalysing the translocation of inorganic cations;<br>nucleoside-triphosphate phosphatase |
| group_2244 | 27 | 1 | 1 | N-acetylmuramoyl-L-alanine amidase | F:GO:0016787<br>P:GO:1901135; | F:hydrolase activity<br>P:carbohydrate derivative metabolic process; | EC:3.5.1.28 | N-acetylmuramoyl-L-alanine amidase |
| group_2498 | 27 | 1 | 1 | glucosaminidase domain-containing protein | F:GO:0016787 | F:hydrolase activity | EC:3.5.1.4 | amidase |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| group_2439 | 27 | 1 | 2 | terminase TerL endonuclease subunit | F:GO:0003824 | F:catalytic activity | | |
| *bioF* | 27 | 1 | 3 | 8-amino-7-oxononanoate synthase | P:GO:0006766; P:GO:0006790; F:GO:0016740 | P:vitamin metabolic process; P:sulfur compound metabolic process; F:transferase activity | EC:2.3.1.47 | 8-amino-7-oxononanoate synthase |
| *bglF* | 27 | 1 | 2 | PTS beta-glucoside transporter subunit IIABC | P:GO:0055085; F:GO:0005215; F:GO:0016740; C:GO:0005886 | P:transmembrane transport; F:transporter activity; F:transferase activity; C:plasma membrane | EC:2.7.1 | Transferring phosphorus-containing groups |
| *yheI_2* | 25 | 0 | 2 | ABC transporter ATP-binding protein | P:GO:0006790; P:GO:0055085; P:GO:0055086; F:GO:0005215; F:GO:0016740; F:GO:0140657 | P:sulfur compound metabolic process; P:transmembrane transport; P:nucleobase-containing small molecule metabolic process; F:transporter activity; F:transferase activity; F:ATP-dependent activity | EC:7.2.2; EC:2.7.1.24 | Catalysing the translocation of inorganic cations; dephospho-CoA kinase |
| group_2412 | 26 | 0 | 2 | acyl-CoA carboxylase subunit beta | P:GO:0006629; F:GO:0016740; F:GO:0016874 | P:lipid metabolic process; F:transferase activity; F:ligase activity | EC:6.4.1.3; EC:2; EC:6.4.1.2 | propionyl-CoA carboxylase; Transferases; acetyl-CoA carboxylase |
| group_2494 | 26 | 0 | 0 | PTS glucose transporter subunit IIA | P:GO:0055085; F:GO:0005215; F:GO:0016740 | P:transmembrane transport; F:transporter activity; F:transferase activity | EC:2.7.1 | Transferring phosphorus-containing groups |
| *rssB_2* | 27 | 1 | 2 | two-component system response regulator YesN[a] | P:GO:0023052 | P:signaling | | |
| *yveA_2* | 27 | 1 | 3 | aspartate/proton symporter AspP | P:GO:0055085; F:GO:0005215 | P:transmembrane transport; F:transporter activity | | |

[1]"SeqName" is the name assigned by the Roary analysis. [2]“The number in Group A” includes the count for the Bekang strain. [3]“Description” refers to the annotation in the sequence similarity search. [a]*rssB* and *yesN* are synonyms for the same gene.

591 **Table 3.** A selection of nearly B-specific genes.

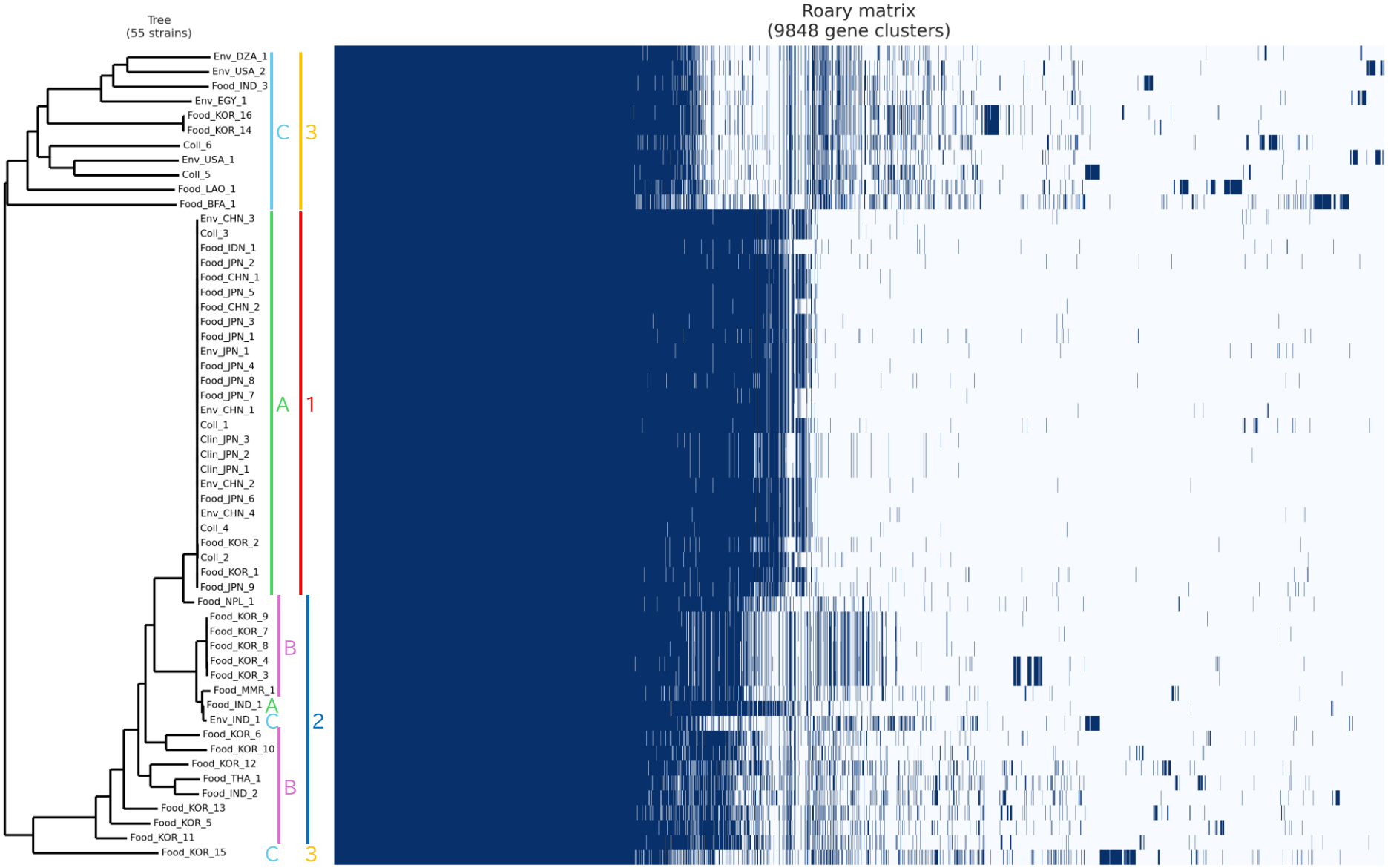
| SeqName[1] | The number in Group A[2] | Presence in Bekang strain (1/0) | The number in Group B | Description[3] | GO IDs | GO Names | Enzyme Codes | Enzyme Names |
|---|---|---|---|---|---|---|---|---|
| *sigV_2* | 0 | 0 | 14 | sigma-70 family RNA polymerase sigma factor | P:GO:0006355; F:GO:0003677; F:GO:0140110 | P:regulation of DNA-templated transcription; F:DNA binding; F:transcription regulator activity | | |
| *yedK* | 0 | 0 | 12 | SOS response-associated peptidase | P:GO:0006281; F:GO:0003677 | P:DNA repair; F:DNA binding | | |
| group_2711 | 0 | 0 | 12 | helix-turn-helix transcriptional regulator | F:GO:0003677 | F:DNA binding | | |
| group_6178 | 4 | 0 | 13 | IS21 family transposase | P:GO:0006310; F:GO:0003677; F:GO:0140097 | P:DNA recombination; F:DNA binding; F:catalytic activity, acting on DNA | | |
| *rapI_4* | 0 | 0 | 13 | tetratricopeptide repeat protein | F:GO:0005515 | F:protein binding | | |
| *rsoA* | 0 | 0 | 12 | sigma-O factor regulator RsoA | F:GO:0005515 | F:protein binding | | |
| group_393 | 3 | 0 | 15 | M20/M25/M40 family metallo-hydrolase | F:GO:0016787 | F:hydrolase activity | EC:3 | Hydrolases |
| group_1244 | 5 | 0 | 12 | ATP-dependent helicase | F:GO:0016787 | F:hydrolase activity | EC:3 | Hydrolases |
| group_2724 | 0 | 0 | 12 | N-acetylmuramoyl-L-alanine amidase | P:GO:1901135; F:GO:0016787 | P:carbohydrate derivative metabolic process; F:hydrolase activity | EC:3.5.1.28 | N-acetylmuramoyl-L-alanine amidase |
| group_2746 | 0 | 0 | 12 | Fe(3+)-citrate ABC transporter ATP-binding protein YfmF | F:GO:0016787; F:GO:0140657 | F:hydrolase activity; F:ATP-dependent activity | EC:3.6.1.15 | nucleoside-triphosphate phosphatase |
| *spoVAD_1* | 0 | 0 | 12 | stage V sporulation protein AD | F:GO:0016740 | F:transferase activity | EC:2.3 | Acyltransferases |
| *uhpB* | 1 | 1 | 12 | sensor histidine kinase | P:GO:0023052; F:GO:0016740; F:GO:0140096; F:GO:0140299 | P:signaling; F:transferase activity; F:catalytic activity, acting on a protein; | EC:2.7.3; EC:2.7.13.3 | Transferring phosphorus-containing groups; histidine kinase |

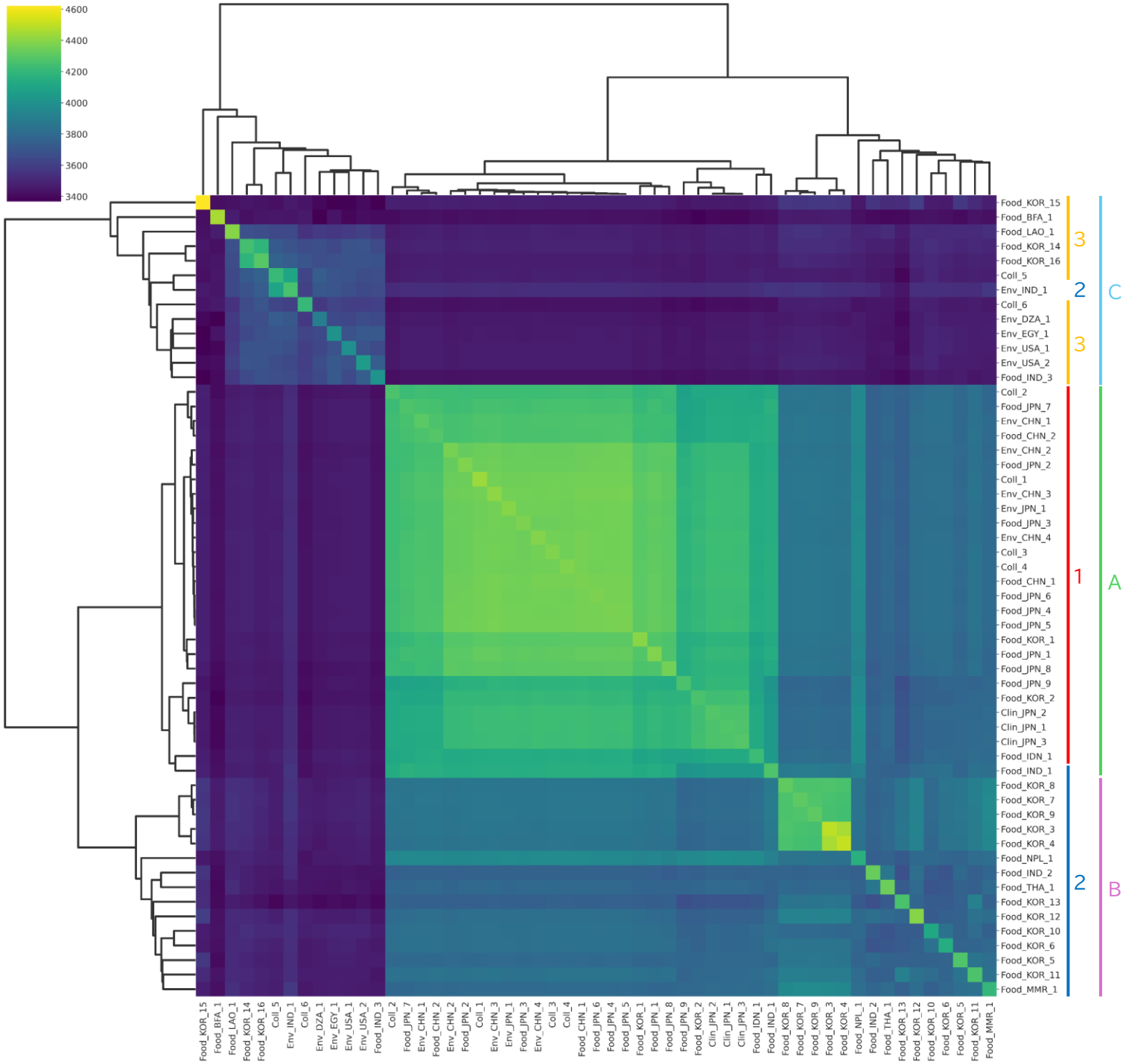
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | F:molecular sensor activity | | |
| *ydaF_2* | 0 | 0 | 13 | GNAT family protein | F:GO:0016740 | F:transferase activity | EC:2.3.1 | Acyltransferases |
| *yoaC* | 0 | 0 | 13 | FGGY-family carbohydrate kinase | P:GO:0005975; F:GO:0016740 | P:carbohydrate metabolic process; F:transferase activity | EC:2.7 | Transferring phosphorus-containing groups |
| group_96 | 2 | 1 | 14 | peptidoglycan O-acetyltransferase OatA | P:GO:0005975; P:GO:0006629; P:GO:1901135; F:GO:0016740; C:GO:0005886 | P:carbohydrate metabolic process; P:lipid metabolic process; P:carbohydrate derivative metabolic process; F:transferase activity; C:plasma membrane | EC:2.3.1 | Acyltransferases |
| group_428 | 2 | 1 | 12 | type III polyketide synthase BpsA | F:GO:0016740 | F:transferase activity | EC:2.3 | Acyltransferases |
| group_2741 | 0 | 0 | 12 | 8-amino-7-oxononanoate synthase | P:GO:0006766; P:GO:0006790; F:GO:0016740 | P:vitamin metabolic process; P:sulfur compound metabolic process; F:transferase activity | EC:2.3.1.47 | 8-amino-7-oxononanoate synthase |
| group_3003 | 0 | 0 | 12 | GNAT family N-acetyltransferase | F:GO:0016740 | F:transferase activity | EC:2.3.1 | Acyltransferases |
| group_2736 | 0 | 0 | 13 | phospholipase D-like domain-containing protein | F:GO:0003824 | F:catalytic activity | | |
| group_3006 | 0 | 0 | 15 | T7SS effector LXG polymorphic toxin | F:GO:0003824 | F:catalytic activity | | |
| group_2720 | 0 | 0 | 12 | 4-hydroxyphenylacetate 3-monooxygenase, oxygenase component | F:GO:0016491 | F:oxidoreductase activity | EC:1.3; EC:1.14.14 | Acting on the CH-CH group of donors; Acting on paired donors, with incorporation or reduction of molecular oxygen. The oxygen incorporated need not be derived from O2 |

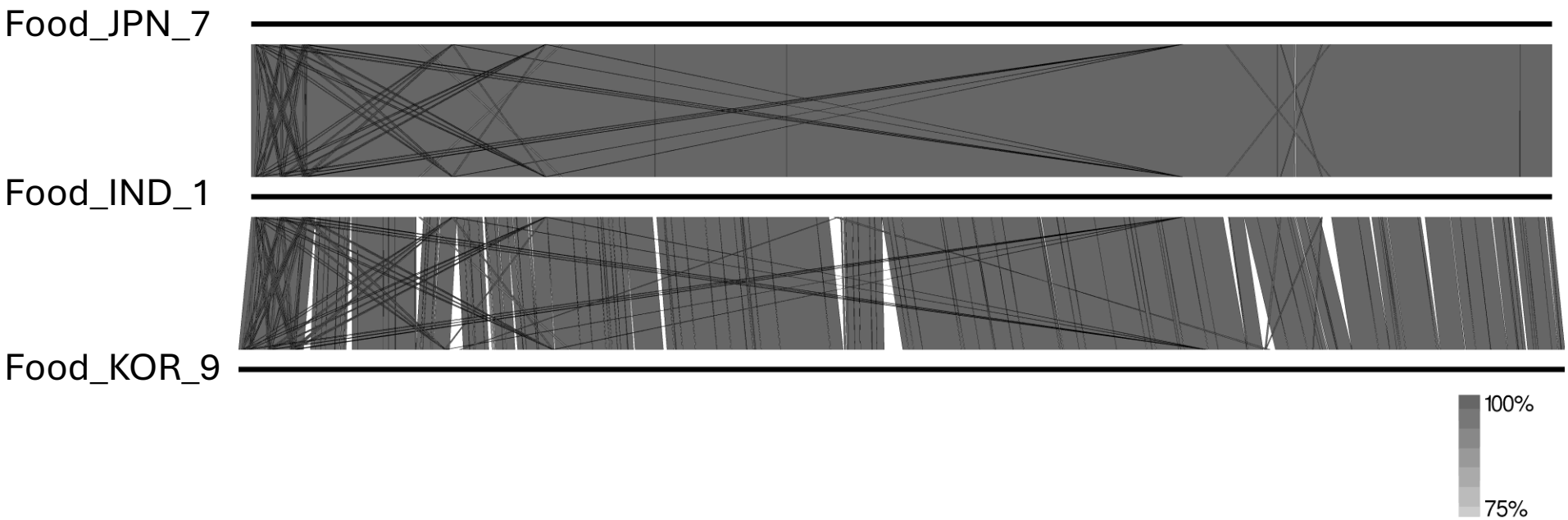
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *alkH* | 0 | 0 | 12 | aldehyde dehydrogenase family protein | F:GO:0016491 | F:oxidoreductase activity | EC:1.2.1 | Acting on the aldehyde or oxo group of donors |
| *yfmJ* | 0 | 0 | 12 | NADP-dependent oxidoreductase | F:GO:0016491 | F:oxidoreductase activity | EC:1.3.1 | Acting on the CH-CH group of donors |
| *pelB_2* | 0 | 0 | 12 | polysaccharide lyase family 1 protein | P:GO:0005975; F:GO:0016829; C:GO:0005576 | P:carbohydrate metabolic process; F:lyase activity; C:extracellular region | EC:4.2.2.2 | pectate lyase |
| *yxxD* | 0 | 0 | 13 | type II toxin-antitoxin system antitoxin YxxD | C:GO:0005737 | C:cytoplasm | | |
| group_2719 | 0 | 0 | 12 | YoaK family protein | C:GO:0005886 | C:plasma membrane | | |
| group_2712 | 0 | 0 | 12 | DUF2975 domain-containing protein | C:GO:0005886 | C:plasma membrane | | |
| group_1989 | 0 | 0 | 12 | DUF421 domain-containing protein | C:GO:0005886 | C:plasma membrane | | |
| group_3077 | 0 | 0 | 12 | RDD family protein | C:GO:0005886 | C:plasma membrane | | |
| group_3076 | 0 | 0 | 12 | DUF5391 family protein | C:GO:0005886 | C:plasma membrane | | |
| group_2664 | 1 | 1 | 15 | DUF2892 domain-containing protein | C:GO:0005886 | C:plasma membrane | | |
| group_2665 | 1 | 1 | 14 | NP_388648.1 conserved membrane protein | C:GO:0005886 | C:plasma membrane | | |
| group_1986 | 0 | 0 | 12 | Uncharacterised protein | C:GO:0016020 | C:membrane | | |
| *yoaH* | 0 | 0 | 12 | methyl-accepting chemotaxis protein | C:GO:0016020 | C:membrane | | |
| group_2716 | 0 | 0 | 12 | DUF4944 domain-containing protein | C:GO:0016020 | C:membrane | | |
| group_3078 | 0 | 0 | 12 | DUF418 domain-containing protein | C:GO:0016020 | C:membrane | | |
| group_2733 | 0 | 0 | 12 | MULTISPECIES: stage V sporulation protein AE | C:GO:0016020 | C:membrane | | |
| group_2731 | 0 | 0 | 12 | stage V sporulation protein AC | C:GO:0016020 | C:membrane | | |
| group_3007 | 0 | 0 | 15 | KFK80766.1 putative membrane protein | C:GO:0016020 | C:membrane | | |
| group_631 | 1 | 1 | 15 | YgcG family protein | C:GO:0016020 | C:membrane | | |
| group_2672 | 1 | 1 | 13 | YpfB family protein | C:GO:0016020 | C:membrane | | |

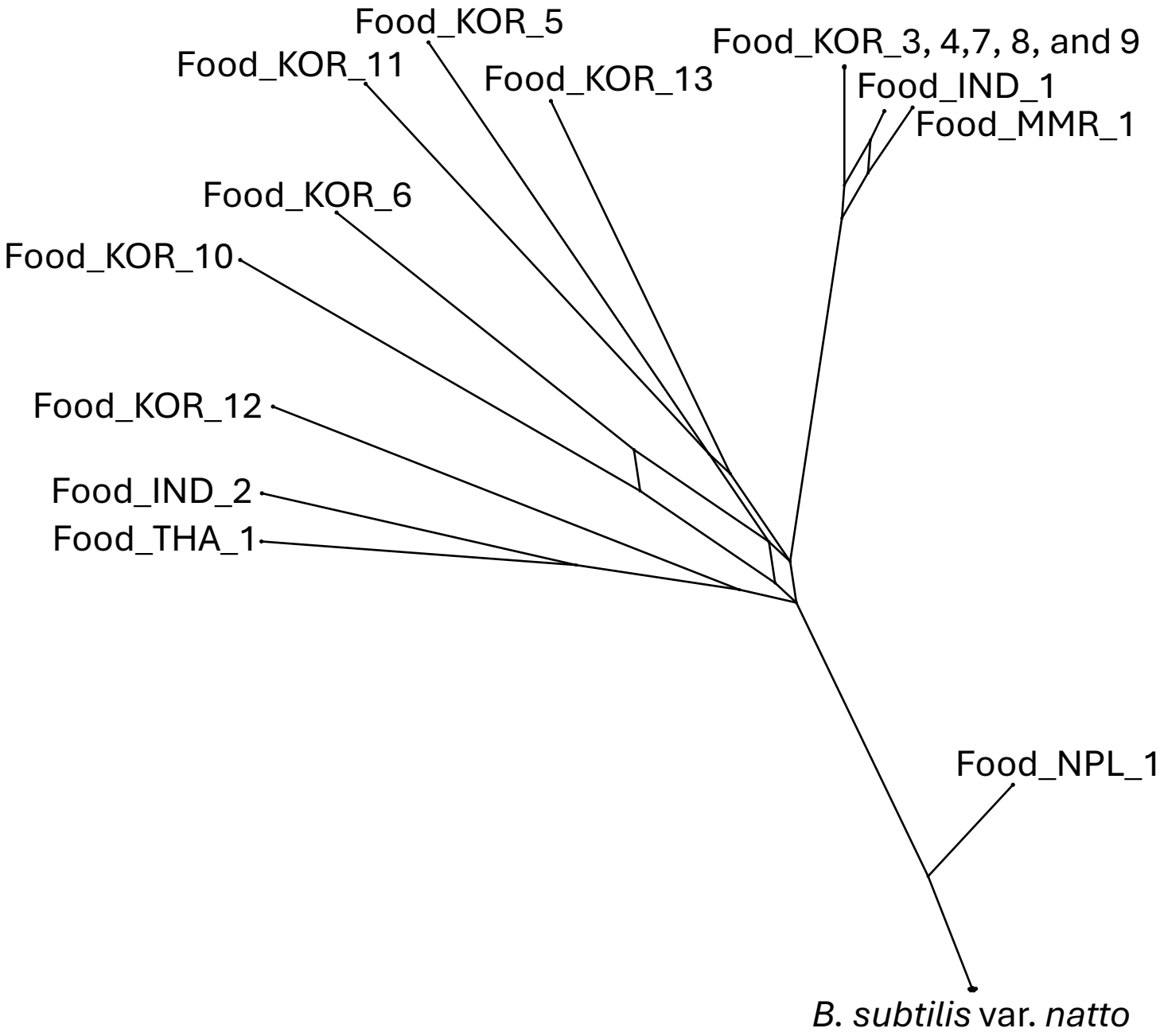
| | | | | | | |
|---|---|---|---|---|---|---|
| group_385 | 4 | 1 | 14 | MFS transporter | C:GO:0016020 | C:membrane |
| *yfiZ_1* | 0 | 0 | 12 | Fe(3+) dicitrate ABC transporter permease FecD | F:GO:0005215 | F:transporter activity |
| *fecD* | 0 | 0 | 12 | Fe(3+)-citrate ABC transporter permease YfmE | F:GO:0005215 | F:transporter activity |
| group_303 | 2 | 1 | 14 | iron ABC transporter permease | F:GO:0005215 | F:transporter activity |
| *gerXA_1* | 0 | 0 | 12 | Spore germination protein | P:GO:0009847; C:GO:0016020 | P:spore germination; C:membrane |
| group_2742 | 0 | 0 | 15 | sugar ABC transporter permease | P:GO:0055085 | P:transmembrane transport |
| group_341 | 0 | 0 | 14 | aspartate/proton symporter AspP | P:GO:0055085; F:GO:0005215 | P:transmembrane transport; F:transporter activity |
| group_1149 | 1 | 1 | 13 | GABA permease | P:GO:0055085; F:GO:0005215 | P:transmembrane transport; F:transporter activity |
| group_5097 | 4 | 0 | 13 | IS21-like element helper ATPase IstB | F:GO:0005524 | F:ATP binding |
| *oxdD* | 0 | 0 | 12 | Oxalate decarboxylase | P:GO:0033609 | P:oxalate metabolic process |

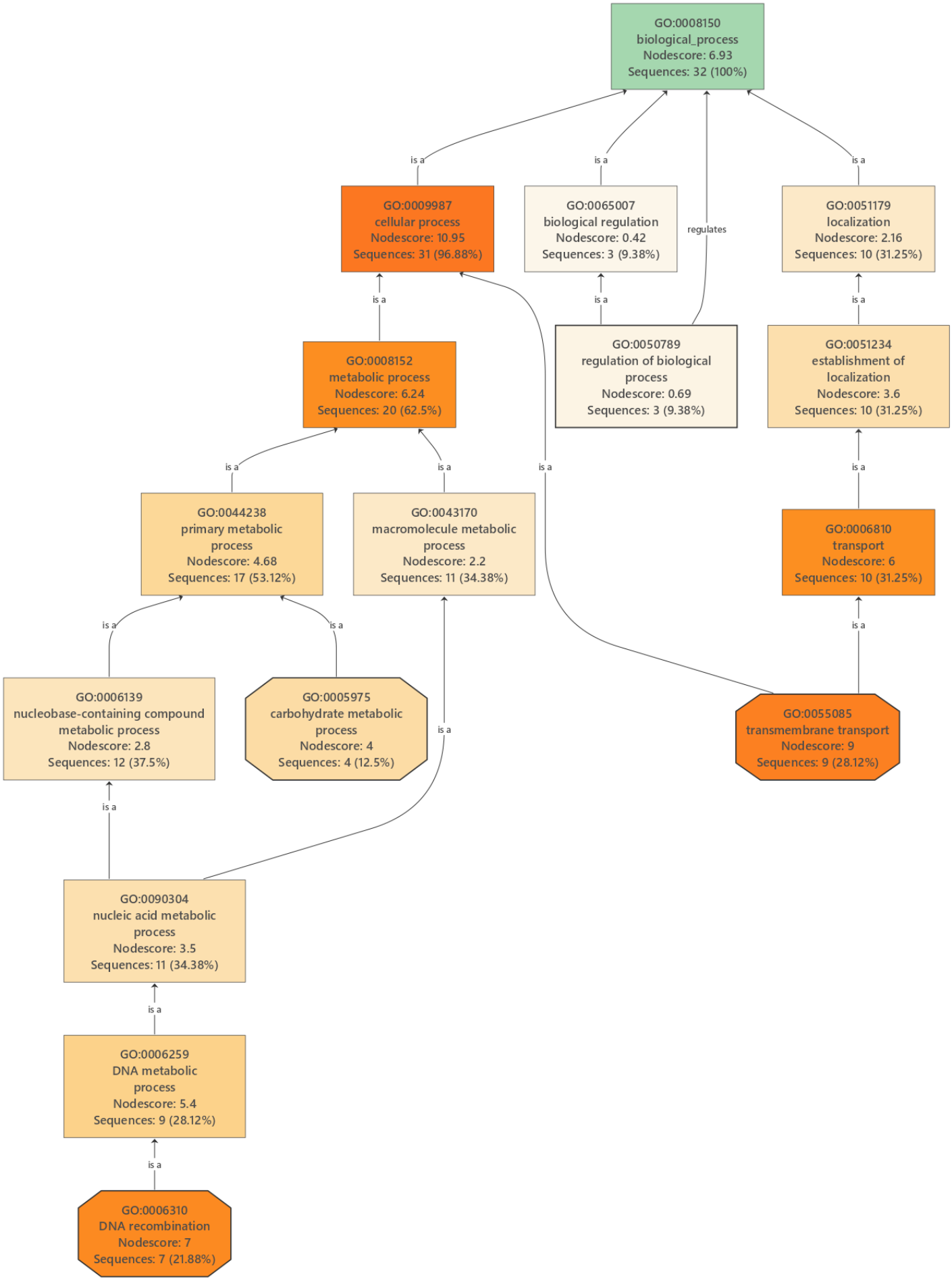
[1]"SeqName" is the name assigned by the Roary analysis. [2]"The number in Group A" includes the count for the Bekang strain. [3]"Description" refers to the annotation in the sequence similarity search.

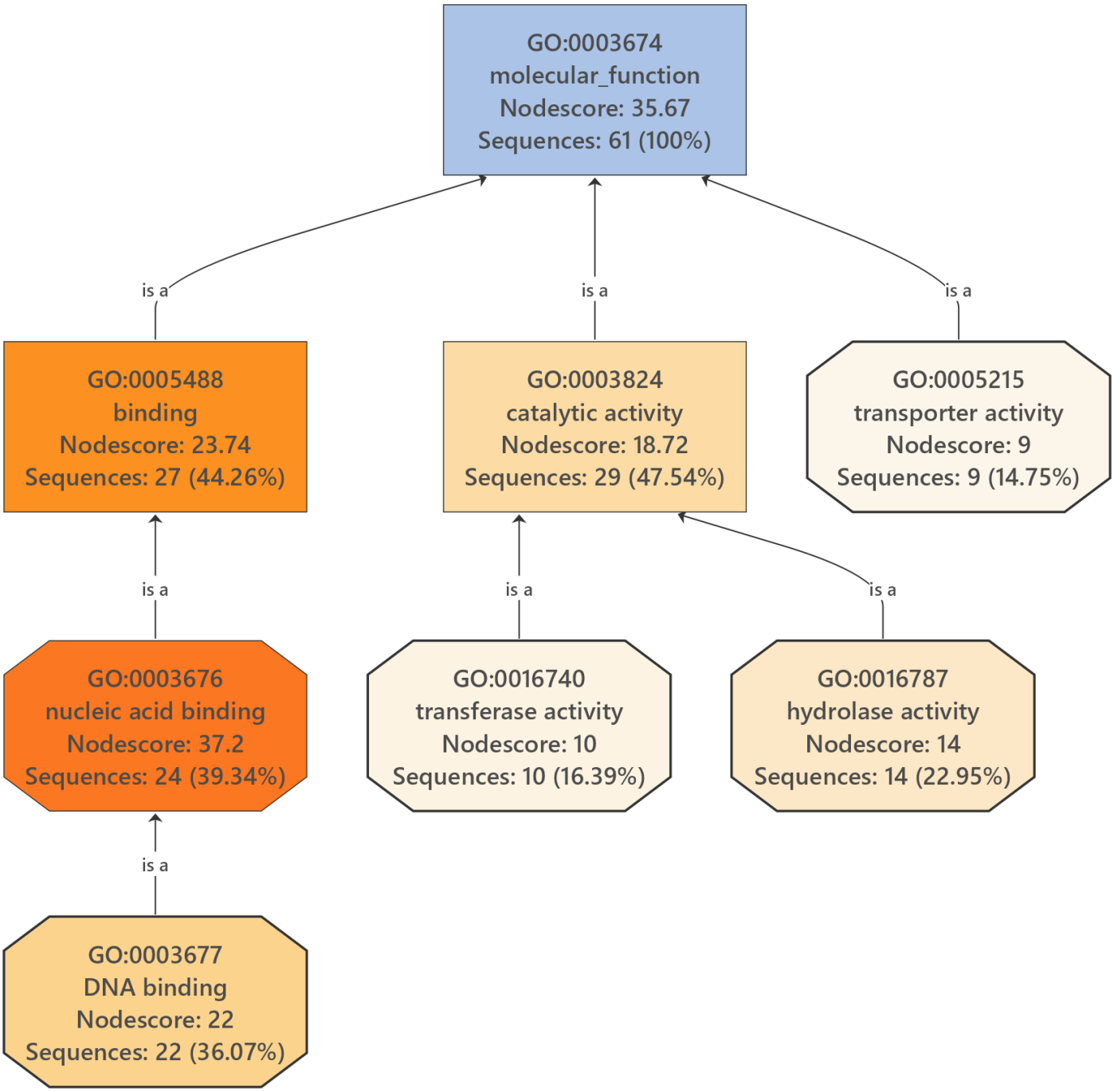

Tree
(55 strains)
Roary matrix
(9848 gene clusters)
Env_DZA_1
Env_USA_2
Food_IND_3
Env_EGY_1
Food_KOR_16
Food_KOR_14
Coll_6
Env_USA_1
Coll_5
Food_LAO_1
Food_BFA_1
Env_CHN_3
Coll_3
Food_IDN_1
Food_JPN_2
Food_CHN_1
Food_JPN_5
Food_CHN_2
Food_JPN_3
Food_JPN_1
Env_JPN_1
Food_JPN_4
Food_JPN_8
Food_JPN_7
Env_CHN_1
Coll_1
Clin_JPN_3
Clin_JPN_2
Clin_JPN_1
Env_CHN_2
Food_JPN_6
Env_CHN_4
Coll_4
Food_KOR_2
Coll_2
Food_KOR_1
Food_JPN_9
Food_NPL_1
Food_KOR_9
Food_KOR_7
Food_KOR_8
Food_KOR_4
Food_KOR_3
Food_MMR_1
Food_IND_1
Env_IND_1
Food_KOR_6
Food_KOR_10
Food_KOR_12
Food_THA_1
Food_IND_2
Food_KOR_13
Food_KOR_5
Food_KOR_11
Food_KOR_15
C
3
A
1
B
A
C
2
B
C
3


4600
4400
4200
4000
3800
3600
3400
Food_KOR_15
Food_BFA_1
Food_LAO_1
Food_KOR_14
Food_KOR_16
Coll_5
Env_IND_1
Coll_6
Env_DZA_1
Env_EGY_1
Env_USA_1
Env_USA_2
Food_IND_3
Coll_2
Food_JPN_7
Env_CHN_1
Food_CHN_2
Env_CHN_2
Food_JPN_2
Coll_1
Env_CHN_3
Env_JPN_1
Food_JPN_3
Env_CHN_4
Coll_3
Coll_4
Food_CHN_1
Food_JPN_6
Food_JPN_4
Food_JPN_5
Food_KOR_1
Food_JPN_1
Food_JPN_8
Food_JPN_9
Food_KOR_2
Clin_JPN_2
Clin_JPN_1
Clin_JPN_3
Food_IDN_1
Food_IND_1
Food_KOR_8
Food_KOR_7
Food_KOR_9
Food_KOR_3
Food_KOR_4
Food_NPL_1
Food_IND_2
Food_THA_1
Food_KOR_13
Food_KOR_12
Food_KOR_10
Food_KOR_6
Food_KOR_5
Food_KOR_11
Food_MMR_1
3
2
3
1
2
C
A
B


Food_JPN_7
Food_IND_1
Food_KOR_9
100%
75%


Food_KOR_5
Food_KOR_11
Food_KOR_13
Food_KOR_3, 4,7, 8, and 9
Food_IND_1
Food_MMR_1
Food_KOR_6
Food_KOR_10
Food_KOR_12
Food_IND_2
Food_THA_1
Food_NPL_1
*B. subtilis* var. *natto*


GO:0008150
biological_process
Nodescore: 6.93
Sequences: 32 (100%)
is a
is a
is a
GO:0009987
cellular process
Nodescore: 10.95
Sequences: 31 (96.88%)
GO:0065007
biological regulation
Nodescore: 0.42
Sequences: 3 (9.38%)
regulates
GO:0051179
localization
Nodescore: 2.16
Sequences: 10 (31.25%)
is a
is a
is a
GO:0008152
metabolic process
Nodescore: 6.24
Sequences: 20 (62.5%)
GO:0050789
regulation of biological process
Nodescore: 0.69
Sequences: 3 (9.38%)
GO:0051234
establishment of localization
Nodescore: 3.6
Sequences: 10 (31.25%)
is a
is a
is a
is a
GO:0044238
primary metabolic process
Nodescore: 4.68
Sequences: 17 (53.12%)
GO:0043170
macromolecule metabolic process
Nodescore: 2.2
Sequences: 11 (34.38%)
GO:0006810
transport
Nodescore: 6
Sequences: 10 (31.25%)
is a
is a
is a
is a
GO:0006139
nucleobase-containing compound metabolic process
Nodescore: 2.8
Sequences: 12 (37.5%)
GO:0005975
carbohydrate metabolic process
Nodescore: 4
Sequences: 4 (12.5%)
GO:0055085
transmembrane transport
Nodescore: 9
Sequences: 9 (28.12%)
is a
is a
GO:0090304
nucleic acid metabolic process
Nodescore: 3.5
Sequences: 11 (34.38%)
is a
GO:0006259
DNA metabolic process
Nodescore: 5.4
Sequences: 9 (28.12%)
is a
GO:0006310
DNA recombination
Nodescore: 7
Sequences: 7 (21.88%)


GO:0003674
molecular_function
Nodescore: 35.67
Sequences: 61 (100%)
is a
is a
is a
GO:0005488
binding
Nodescore: 23.74
Sequences: 27 (44.26%)
GO:0003824
catalytic activity
Nodescore: 18.72
Sequences: 29 (47.54%)
GO:0005215
transporter activity
Nodescore: 9
Sequences: 9 (14.75%)
is a
is a
is a
GO:0003676
nucleic acid binding
Nodescore: 37.2
Sequences: 24 (39.34%)
GO:0016740
transferase activity
Nodescore: 10
Sequences: 10 (16.39%)
GO:0016787
hydrolase activity
Nodescore: 14
Sequences: 14 (22.95%)
is a
GO:0003677
DNA binding
Nodescore: 22
Sequences: 22 (36.07%)