Exploring Probiotic Potential: A Comparative Genomics and In Silico Assessment of Genes within the Genus *Geobacillus*

Ishfaq Nabi Najar[1], Prayatna Sharma[2], Rohit Das[3], Krishnendu Mondal[4], Ashish Kumar Singh[5], Anu Radha[6], Varsha Sharma[7], Sonali Sharma[8], Nagendra Thakur[9], Sumit G. Gandhi[10] and Vinod Kumar[11]*

Authors:

Dr. Ishfaq Nabi Najar[1], Research Associate, Fermentation and Microbial Biotechnology Division, CSIR IIIM, Jammu India. Email: urooj.ishfaq@gmail.com.

Prayatna Sharma[2], Ph.D. Scholar, Department of Microbiology, School of Life Sciences, Sikkim University, 6th Mile, Samdur, Tadong, Gangtok – 737102, Sikkim, India. Email: psharma.su@gmail.com.

Rohit Das[3], Ph.D. Scholar, Department of Microbiology, School of Life Sciences, Sikkim University, 6th Mile, Samdur, Tadong, Gangtok – 737102, Sikkim, India. Email: rohitd322@gmail.com.

Krishnendu Mondal[4], Ph.D. Scholar, Department of Microbiology, Vidyasagar University, Midnapur India. Email: kmondals99@gmail.com.

Ashish Kumar Singh[5], Research Associate, Department of Biotechnology and Synthetic Biology, CIAB, Mohali, India. Email: a.singhmicro@gmail.com.

Anu Radha[6], Ph.D. Scholar, Fermentation and Microbial Biotechnology Division, CSIR IIIM, Jammu India. Email: anur8492@gmail.com.

Varsha Sharma[7], Ph.D. Scholar, Fermentation and Microbial Biotechnology Division, CSIR IIIM, Jammu India. Email: sophiasharma224@gmail.com.

Sonali Sharma[8], MSc, Fermentation and Microbial Biotechnology Division, CSIR IIIM, Jammu India. Email: sonalisonali607@gmail.com.

Dr. Nagendra Thakur[9]; Associate Professor, Department of Microbiology, School of Life Sciences, Sikkim University, 6th Mile, Samdur, Tadong, Gangtok – 737102, Sikkim, India. Email: nthakur@cus.ac.in.

Dr. Sumit G. Gandhi[10]; Sr. Principal Scientist, Infectious Diseases Division. CSIR IIIM Jammu, India. Email: sumit@iiim.res.in.

*Corresponding Author: Dr. Vinod Kumar[11], Scientist, Fermentation and Microbial Biotechnology Division, CSIR IIIM, Jammu India. Phone No. +91-8427595494, Email: vinod.udsc@iiim.res.in.

Abstract

The pursuit of new probiotic targets has seen a surge, aided by next-generation sequencing, facilitating a thorough exploration of bacterial traits. The genus *Geobacillus* stands out as a promising target for uncovering its potential as a probiotic. The study explored the genetic attributes of the genus *Geobacillus* for their resilience to gastrointestinal conditions, nutrient production, and immunomodulatory compound creation, revealing potential probiotic traits. Additionally, the research undertook predictive analyses of genomic elements such as prophages, CRISPR-Cas systems, insertion sequences, genomic islands, antibiotic resistance genes, and CAZymes. These evaluations aimed to assess the safety aspects associated with the genus *Geobacillus*. A comparative genomic analysis was also carried out using 18 validly published genomes of the genus *Geobacillus* and a few other genomes of *Lactobacillus* and *Bifidobacterium* were taken as control. Genes associated with probiotic traits like adhesion, stress tolerance (acid/bile, osmotic, oxidative), immune modulation, and molecular chaperones were uniformly detected in the *Geobacillus* genus. Notably, mobile genetic elements such as plasmids, prophages, and insertion sequences were absent, as were virulence factors, toxins, and Antibiotic resistance genes. Additionally, CRISPR-Cas systems and CAZymes were present. The pan-genome encompassed 25,284 protein-coding genes with translation. Comparative genomic analysis revealed an open pan-genome for *Geobacillus*. Pan-genome exhibited variability, particularly in genes linked to environmental interaction and secondary metabolite synthesis. In conclusion, *Geobacillus* appears potentially safe and well-suited for the gut habitat. However, further *in vitro* studies are essential to add to the knowledge of the probiotic potential of *Geobacillus* species.

Importance

This comprehensive study highlights the significant probiotic potential and genetic makeup of the *Geobacillus* genus, shedding light on its unique attributes in adapting to extreme environmental conditions. Understanding the probiotic properties of *Geobacillus* is crucial amidst growing concerns over antibiotic resistance, offering promising alternatives for combating pathogenic microbes. Additionally, exploring the genetic diversity and adaptive mechanisms of *Geobacillus* through genomic and metagenomic approaches provides valuable insights into its biotechnological applications and evolutionary history. By employing in-silico methods and comparative analyses with established probiotic genera, this study contributes to elucidating the probiotic characteristics

of *Geobacillus*, paving the way for further research in harnessing its beneficial traits for various applications in health, biotechnology, and environmental remediation.

Key Words: *Geobacillus*, Probiotics, Pan-Genome, Comparative Genomics, Probiogenomics, Phylogenetics

Introduction

The Earth encompasses unique repositories characterized by extreme environmental conditions, where some microbes termed extremophiles thrive with the aid of evolutionary adaptations required for survival (1). Particularly, the genus *Geobacillus* belongs to obligate thermophilic, Gram-positive rods, aerobic or facultative aerobic species, thriving in the warmest regions on earth, from geothermal springs to equatorial deserts. Surviving in regions accustomed to 45-80°C, these obligatory thermophiles possess optimum growth at 55-65°C (2, 3). Previously known to belong to the genus *Bacillus*, the reclassified genus *Geobacillus* holds immense biotechnological importance. Their habitat remains remarkably diverse, ranging from cold environments such as Antarctica (4) to extremely hot niches (5). Their ability to withstand harsh temperature conditions paves their way towards numerous biotechnological applications; these species serve as key hosts for thermostable proteins with robust catalytic activity, such as amylases, protease, and lipases; and also serve various roles in bioprocessing applications such as bioproduction of biofuels and biodiesel and bioremediation (2, 6).

The genus *Geobacillus* was previously classified as a separate unit, known as group 5 of the thermophilic genus *Bacillus*, as demonstrated by a 16S rRNA gene sequencing study by Ash et al, 1991 (7). However, due to genetic heterogeneity, the revision of taxonomic classification let the group be reclassified as *Geobacillus* (7). Currently, there are 20 species prescribed under the genus *Geobacillus* (8). Research since the last century has involved the reclassification of the genus *Bacillus* to *Geobacillus*, with the earliest report of isolation of *Bacillus stearothermophilus* (now termed *Geobacillus stearothermophilus*) by Donk (9). Until the end of the 21st century, a taxonomic study carried out by (10) revealed species of *B. kaustophilus*, *B. stearothermophilus*, *B.*

thermocatenulatus, *B. thermodenitrificans*, *B. thermoglucosidasius*, and *B. thermoleovorans* exhibited similar physiological and phylogenetic properties, thereby classifying these species into a new genus "*Geobacillus*". With most of the species with a Gram-positive nature and few showing a Gram-variable nature, the bacteria appear in short chains and may or may not be motile. With an optimum temperature for growth at 55–65 °C, these species are obligate thermophiles and exhibit growth in neutral pH ranges (6.0–8.5) (1). The nutritional profile of *Geobacillus* emerges as chemoorganotrophs, with oxygen as the terminal electron acceptor (with few cases showing nitrate as an electron acceptor). The species require no special nutrients like vitamins and can be grown in simple media such as nutrient agar, with basic requirements like carbohydrates, peptone, and tryptone as primary carbon and energy sources (11). Various studies have been done to evaluate the biotechnological potential of the *Geobacillus* species, however, there are very few studies done on the probiotic potential of the genus *Geobacillus*.

Probiotics that are often considered live microbes confer health benefits when administered due to their antagonistic activity against their surrounding microbes (12). Considering the immense global threat that antibiotic resistance poses to humanity today (13), there remains a constant need for novel strains with enhanced probiotic properties that confer stronger activity against pathogenic microbes. Members belonging to the *Firmicutes* have gained considerable attention for their probiotic properties for humans and animals (14, 15). The genera *Geobacillus* and *Parageobacillus* secrete anti-microbial compounds such as bacteriocins, antibiotics, and bacteriocin-like inhibitory substances (BLISes) and serve as potential candidates exhibiting unique probiotic properties (12). For instance, a preliminary study on *Geobacillus thermoleovorans* isolated from oil waste showed antimicrobial activity against pathogenic strains of *Salmonella typhimurium*, *Vibrio parahemolitycu*s, *Staphylococcus aureus* as evident by antagonism and adherence assay, denoting the strain as a potential probiotic candidate (16). The ability to adhere to abiotic surfaces to create a bio-secure environment may assist in inhibiting the growth of pathogenic strains (12). Other mechanisms include the inhibition of Quorum-sensing in Gram-negative bacteria by *Geobacillus* strains, due to their Acyl Homoserine Lactones (AHLs) degrading ability (17). The plausible explanation for their candidature in the ever-increasing probiotic market may be due to their enhanced survival strategies as compared to the routinely used probiotic strains. Additionally, these organisms are part of a group that is both biologically safe and non-pathogenic. They are recognized as a bimodal probiotic microbiota present in both humans and animals and are an

essential composite of the soil, skin, gastrointestinal, and plant microbiome (15). The strains of *Geobacillus* thrive in extreme environmental parameters of high temperature, different pH ranges, UV radiation, heavy metals, salts, organic/inorganic chemicals and detergents, and arid and minuscule moisture environments, mainly owned by their ability to form durable endospores (12, 15).

Various reports suggest the prevalence of antibiotic resistance in mesophilic bacterial species, however, the phenomenon of antibiotic resistance in thermophiles is lesser known. It remains crucial to unravel the antibiotic resistance patterns and the intricate mechanism involved in thermophilic bacteria, given the immense biotechnological applications that the thermophilic bacteria impose (Najar et al., 2020c). Understanding antibiotic resistance in thermophiles remains crucial. Very few studies suggest the study of antibiotic-resistance profiling among thermophilic bacteria. Given this, Najar et al, 2022 conducted research to study antibiotic resistance prevailing in the microbiota of Sikkim hot springs. Both the culture-dependent and PCR amplification analyses revealed the absence of antibiotic resistance by the thermophilic bacteria. Similarly, no antibiotic-resistance genes were found in *Geobacillus* species, suggesting that the genus susceptible to antibiotics natively (13). However, it is crucial to evaluate the genetic makeup of these *Geobacillus* species, particularly emphasizing the diverse attributes that contribute to their probiotic potential using metagenomic and pan-genomic approaches.

One of the approaches such as standard Sanger sequencing is well known and over the years, numerous sequencing platforms like the next generation sequencing have emerged that employ different sequencing strategies to unravel the genomic and metabolic potential of the genome (19). Genome sequencing infers to unraveling the sequences of the entire genome of an organism, instead of sequencing individual genes (20). The whole genome shotgun sequencing method helps in understanding the metabolic and genomic potential of the repertoire of genes present in an organism. Deciphering the complete genome of an organism helps to provide insight into understanding the metabolic potential, metabolomics, functional genes, and adaptive mechanisms (21). For instance, the whole genome sequence of a putative novel strain of thermophilic bacterium *Parageobacillus* sp. indicates the presence of various important genes for carbohydrate, Sulphur, nitrogen, and phosphorus metabolism (21). Another widely employed method is the metagenomic approach which helps in elucidating the diversity of microbes present and crucial genes that are

175 responsible for conferring adaptive strategies to the microbiome, such as antibiotic resistance or
176 stress-related genes. Functional metagenomic study on hot spring samples of Sikkim hot springs
177 reveal diverse set of genes required to cope with stressful conditions such as heat shock, acid, and
178 osmotic stress; various multidrug efflux pumps and transport systems, genes for degradation of
179 xenobiotics, etc. (22).

180 Pan-genome analyses are a powerful tool in bacterial taxonomy that may help redefine the
181 categorization of species or, in defining new species (23). The term pan-genome represents the
182 entire set of genes present in a species and is categorized into three main categories- a core genome
183 that comprises sequences shared between all the species; an accessory genome that represents
184 genes shared in some species; and a singleton or unique genome that is present in a particular
185 species only. The core genome comprises genes required for basic metabolic activities such as
186 genes for antibiotic resistance and housekeeping genes. These genes that are conserved usually
187 infer evolutionary relationships amongst different strains. The genes present in the accessory
188 genome participate in adaptive strategies in a strain, where horizontal gene transfer (thus are
189 subjected to gene gain/loss) plays a major role that aids in coping up adaptive mechanisms in a
190 bacterium when subjected to a new niche (24). Pan-genome analysis plays a major role in
191 implicating the evolution of the genus *Geobacillus* from *Bacillus.* There remains limited study
192 about pan-genome analyses of the *Geobacillus* genus, although reports for other bacteria from the
193 neighboring genus *Bacillus* have been reported (25). Bezuidt et al, 2016, provided insights into the
194 role of horizontal gene transfer in the diversification of *Geobacillus* (26).

195 This study employs an in-silico approach to assess the potential probiotic characteristics of the
196 Geobacillus genus. It delves into various attributes, including taxonomy, phylogenomics,
197 probiotic-related genes, prediction of antibiotic resistance genes, virulence factors, toxins, mobile
198 genetic elements, plasmids, bacteriocins, CRISPR-Cas systems, and CAZymes. Additionally, a
199 pan-genome analysis was conducted. To benchmark and contextualize the findings, comparative
200 analysis involved two established genera, *Bifidobacterium* and *Lactobacillus*, serving as positive
201 controls. Overall, this comprehensive study serves as a significant exploration of both the probiotic
202 potential and genetic makeup of the *Geobacillus* genus.

203

204

Results

Source and quality check of various genomes

Among 18 *Geobacillus* species used in this study, 13 are validly published under ICNP as per LPSN and among 13 *Geobacillus* species used, 7 are type species such as *Geobacillus icigianus* G1w1, *Geobacillus kaustophilus* NBRC 102445, *Geobacillus lituanicus* N-3, *Geobacillus proteiniphilus* 1017, *Geobacillus stearothermophilus* ATCC 12980, *Geobacillus thermocatenulatus* BGSC 93A1, and *Geobacillus thermodenitrificans* DSM 465. In the case of control, probiotic species used are validly published and type species. The quality of genomes was assessed and it was shown that all the genomes possess good quality with >97% completeness and very low fractional contamination as shown in **Table.1. and Supplementary Report File.1.** The GC content of *Geobacillus* genomes range between 49-53%, *Lactobacillus* genomes range between 34-49% and to that of *Bifidobacterium* genomes range between 58-60%. The total genome length of the *Geobacillus* genomes was higher (3.45mb-3.80mb) than those of *Lactobacillus* (2.03mb-3.2mb), and *Bifidobacterium* genomes (2.23mb-2.83mb). The isolation source of most of the *Geobacillus* species analyzed in this study is environmental such as Oil fields, Hot springs, fumaroles, and sediments. However, the isolation source of Geobacillus *stearothermophilus* ATCC 12980 and *Geobacillus thermodenitrificans* DSM 465 was food products such as spoiled canned food and sugar beet juice respectively. On the other hand, the isolation source of other probiotic species was food and gastrointestinal tract as shown in **Table 1.**

Taxonomy and Phylogenomics Analysis

The ANI values were plotted on a heat map **Fig.1.** It has been shown that the *Geobacillus* genomes possess ANI values between 90-100%, thus predicting their high nucleotide identity. Whereas the *Lactobacillus* and *Bifidobacterium* genomes possess ANI values between 60-70%, also concerning genomes of genus *Geobacillus*, thus indicating their low nucleotide identity. The cluster analysis represented by the dendrogram shown in the heat map formed two major clades bifurcating the genus *Geobacillus* and two other genera studied i.e. *Lactobacillus* and *Bifidobacterium* based on ANI values. This supports the above assumption indicating that genomes of genus *Geobacillus* possess high nucleotide identity. The genus *Geobacillus* clade further differentiates into two main sub-clades. It was interesting to see that

the earlier isolated and type species such as *G. stearothermophilus*, *G. icigianus*, *G. thermodenitrificans*, *G. lituanicus*, *G. uzenensis*, and *G. subterraneus* possess little variation in ANI values than the recently isolated genomes which are not validly published. The phylogenetic analysis among the genomes of only the genus *Geobacillus* shows a similar pattern with two major clades and three separate clades have been formed by *G. icigianus, G. jurassicus*, and *G. vulcani* **Fig.2.**

We were also trying to find out the phylogenetic relation among the genus *Geobacillus* and two other probiotic species of genus *Lactobacillus* and *Bifidobacterium* based on COG-based core genes **Fig.3. and Supplementary Fig.1.** and single copy genes using single-copy BV-BRC PGFams. The phylogenetic tree, established using a set of 49 core universal genes defined by COG, indicates the emergence of two primary clades encompassing the genus *Geobacillus* as well as the genus *Lactobacillus* and *Bifidobacterium*. The genus *Geobacillus* further differentiates into two sub clades, whereas species *G. jurassicus*, *G. vulcani*, and *G. icigianus form* separate clades similar to the phylogenetic tree of only genus *Geobacillus* based on ANI values. The genus *Lactobacillus* shows a close relation to *Parageobacillus* species whereas *Lactobacillus* and *Bifidobacterium* are distantly related to genus *Geobacillus* as they possess separate clades. Similar results were shown by phylogenetic tree based on single copy genes as shown in **Fig.4.** However, based on these genes *Lactobacillus* forms a clade close to *G. icigianus*, and other earlier isolated genomes such as *G. stearothermophilus*, *G. thermodenitrificans*, *G. lituanicus*, and *G. subterraneus.* Thus, based on core gene phylogenetic analysis of *Geobacillus, Lactobacillus,* and *Bifidobacterium* genomes, it has been shown that they significantly differ which is further supported by phylogenetic analysis based on single gene analysis.

All the genomes were annotated with Prokka using the KBase platform. However, the cross-genus comparisons were also done using the new Comparative Systems service at BV-BRC which annotates the genomes based on global protein families, known as PGFams using RASTtk4. The functional assessment (subsystem analysis) among the genomes using canonical gene family assignment from COG, PFAM, TIGRFAM, and The SEED has been done using **View Function Profile for Genomes - v1.4.0** and is represented in Heatmap **Fig.5.** it has been shown that subsystems like secondary metabolites biosynthesis, transport,

catabolism, intracellular trafficking, secretion, chromatin structure and dynamics, cytoskeleton, and RNA processing and modification were uncharacterized in all the genomes. However, there is not much difference in other characterized subsystems between genomes of the genus *Geobacillus*, *Lactobacillus*, and *Bifidobacterium*. Cell motility is absent in the genus *Lactobacillus* and *Bifidobacterium* whereas present in the genus *Geobacillus*. Moreover, there are slight differences in a few subsystems among the studied genus such as Energy production and conversion, and lipid transport and metabolism. This signifies that there might be similarities among these genera considering probiotic properties.

Identification of Genes Related to Probiotics

Genes involved in the mechanism of modulation of the immune system, vitamin biosynthesis, fatty acid synthesis, resistance to stress conditions (acid, bile, osmotic, and oxidative stress), adhesion, bacterial colonization, and molecular chaperones were detected. The predicted genes among all these categories are present differently in **Table 2.** 25 genes of immune system modulation were screened and genes such as *Pyrc*, *PyrE*, *PyrH*, *PyrB*, *PyrG*, *PyrK*, *TxrA*, *TxrB*, *Clpb*, *CysC*, *nrdR* are present in all the genomes whereas *CysE*, *YjbK*, and *YjbM* were present only in *Geobacillus* genomes and *nrdH* and *dltC* are present only in *Bifidobacterium* and *Lactobacillus* genomes. 25 genes of adhesion were screened and genes such as *Tuf*, *clpB*, *clpX*, and *clpC* are present in all the genomes. *comC*, *srtD*, *PilT*, *PilZ*, and *lapA* are present in only the *Geobacillus* genomes whereas *dltD* and *dltA* were present in *Bifidobacterium* and *Lactobacillus* genomes only. Three sortase genes such as *strA1*, *strA2*, and *strA3* were absent in all genomes except *srtA1* was present in *Lactobacillus* genomes only. Also, *PilA*, *PilB*, *PilC*, *PilO*, and *PilM* adhesion genes were absent in all the genomes. Six bacterial colonization genes were screened, and only gene *TadA* was present in all the genomes rest were absent such as *TadB*, *TadC*, *TadE*, *TadF*, and *TadZ*. All the acid stress-related genes such as *atpB*, *atpC*, *atpD*, *atpE*, *atpF*, *atpG*, *atpH*, *recA*, *sodA*, *luxS*, *glmU*, *glmS*, *glmM1*, *glmM2* and asps were present in all the genomes whereas *cspA* and *gpmA* were only present in *Bifidobacterium* and *Lactobacillus* genomes. Some acid resistance genes such as *gpmL*, *bshA*, *bshB*, and *bshC* were only present in *Geobacillus* genomes. 8 bile resistance genes such as *nagB*, *pyrG*, *argS*, *rpsC*, *rpsE*, *rplD*, *rplE*, and *rplF* were present in all the genomes. Osmotic stress-related genes were absent in *Bifidobacterium* and *Lactobacillus* genomes, however, *opuD* and *opuC* were present

in some *Geobacillus* genomes, and *opuCA*, *opuCB*, and *opuCC* were only present in *Geobacillus* genomes. Molecular chaperones such as *dnaK*, *dnaJ*, and *dnaG* were present in all the genomes whereas *groES* and *groEL* were present in some *Geobacillus* genomes only. 18 genes related to oxidative stress were screened and among them *taxA*. *TaxB*, *feoB*, and *msrC* were present in all the genomes whereas *ndhH*, *ndhB*, and *ndhC* were present only in *Geobacillus* genomes and *msrB* only in *Lactobacillus*. 7 genes related to vitamin biosynthesis were screened and genes *BtuD*, *copA*, and *copZ* were present in all the genomes whereas *BtuF* and *CsoR* were only present in *Geobacillus* genomes. Principle component analysis was done based on all the parameters (genes) related to probiotics including antibiotic resistance genes. The Principle component 1 (PC1) significantly shows the positive correlation between *Bifidobacteria* and *Lactobacillus* genomes and their relatively less positive correlation with the genomes of *Geobacillus* as shown in **Fig.6.**

Assessment of Antibiotic Resistance Genes, Virulence Factors, and Toxins

The presence of antibiotic resistance genes was anticipated within the genomes under investigation. It was shown that only a few antibiotic-resistance genes such as *vanY*, *vanT*, and *rpoB* were present in all the genomes. However, other genes like *ImrD*, *rpsL*, and *qacJ* were present in *Lactobacillus* genomes as shown in **Table 3.** None of the toxin-related genes were present in *Geobacillus*, *Lactobacillus*, and *Bifidobacterium* genomes. Only a few toxin-like proteins such as *PhoH*, *MazF*, *HicA*, and *PemK* genes were predicted in some *Geobacillus* genomes. However, many genes coding toxins such as *FitB*, *PemK*, *MazF*, *ParB*, *RelB*, *RelE*, *MraZ*, *YoeB*, *PhoH*, and *HipA* toxin-related genes were present in genus *Bifidobacterium* and *Lactobacillus* species as shown in **Table.4.** No virulence factors were found in any genomes studied as shown in **Table.3.**

Assessment of Mobile Genetic Elements, Insertion Sequences, Plasmids, Prophages, Bacteriocins, and CRISPR-Cas Systems

Mobile genetic elements were found to be absent in all genomes except a few mobile elements present in *B. brevis* such as IS/Tn. Putative insertion sequences were present in many *Geobacillus*, *Lactobacillus*, and *Bifidobacterium* genomes such as Putative IME, Putative ICE with T4SS, and Putative IME without identified DR as shown in **Supplementary Table. 1.** Plasmids were found to be absent in all the genomes **Table.3.** Many bacteriocins such as

Circularin_A, ComX1, ComX4, Salivaricin_D, Sactipeptides, Pumilarin, and Geobacillin_I_like were present in various genomes of *Geobacillus* like *G. icigianus*, *G. kaustophilus*, *G. lituanicus*, *G. C56T3*, *G. stearothermophilus*, *G. thermocatenulatus*, *G. vulcani*, etc. However, among the two *Bifidobacterium* genomes, only *B. longum* possesses a few bacteriocins such as Geobacillin_I_like and Propionicin_SM1. Moreover, *Lactobacillus* genomes possess different bacteriocins than that of *Geobacillus* and *Bifidobacterium* genomes such as Plantaricin_E/F/A/N/J/K, Enterocin_X, Acidocin_J1132, Enterolysin_A, and Bacteriocin_helveticin_J as shown in **Table.5**. Among 18 *Geobacillus* genomes, 11 possess prophages mainly within the family of Myoviridae and Siphoviridae. Similarly, *Bifidobacterium* and *Lactobacillus* possess prophages mainly from the family of Siphoviridae shown in **Table. 6**. CRISPR-Cas Systems were found to be present only in *Geobacillus* genomes. *Cas1*, *Cas2*, *Cas3*, *Cas4*, *Cas5h*, *Cas6*, *Cas9*, *Cmr1*, *Cmr3*, *Cmr4*, *Cmr5*, and *Cmr6* are present in many *Geobacillus* genomes such as *G. icigianus*, *G. thermoleovorans*, *G. uzenensis*, *G. stearothermophilus*, *G. subterraneus*, and *G. jurassicus* etc. However, CRISPR-Cas Systems were found to be absent in *Bifidobacterium* and *Lactobacillus* genomes as shown in **Table. 2**.

Identification of Carbohydrate-active Enzymes (CAZyme)

Carbohydrate-active Enzymes (CAZyme) were found to be present in almost all genomes. Glycoside hydrolases were present in abundance and various glycoside hydrolase classes such as GH1, GH13, GH13_1, GH13_2, GH13_4, GH13_4, GH13_10, GH13_12, GH13_13, GH13_14, GH13_20, GH13_21, GH13_23, GH13_29, GH13_31, GH13_36, GH13_37, GH13_39, GH13_41, and GH18 were present with more than 2 hits or genes present. Among Glycosyl transferases GT2, GT4, GT51, and GT108 were present with more than 2 hits. CBM34 and CBM68 are present with more than 2 hits. CE4 and CE9 were the carbohydrate esterases present with more than 2 hits. Polysaccharide lyases were absent in all the genomes. AA1 and AA1_3 were CAZymes with Auxiliary activities having more than 2 hits. The above-mentioned CAZymes (except Polysaccharide lyases) were present in all the genomes as shown in Heatmap **Fig.7.**

Pangenome Analysis

The total number of genomes studied in the pangenome was 23 and among them 18 were *Geobacillus* species, 3 were *Lactobacillus* species and 2 were *Bifidobacterium* species. The pan-genome analysis was done within the genus *Geobacillus* alone and with other genera like *Lactobacillus* and *Bifidobacterium*. The pan-genome analysis of all 23 genomes represented 38393 protein-coding genes (with translation), 30702 genes in homolog families, and 7691 genes in singleton families. The total number of families was 11637, 3764 were homolog families, and 7873 were singleton families. When taking the pan-genome analysis of only the genus *Geobacillus* into consideration, the pan-genome represented 25284 protein-coding genes (with translation), 23504 genes in homolog families, and 1680 genes in singleton families. The total number of families was 4641, 2813 were homolog families and 1828 were singleton families **Supplementary Table 2**. The core pangenome rarefaction curve of the genus *Geobacillus* shows the increasing number of gene clusters with every added genome. Thus, this shows that the *Geobacillus* pan-genome is an open pangenome as shown in **Fig.8**. The pangenome circle plot was contrived to plot the core, non-core, and singletons. The circle plot of the genus *Geobacillus* shows larger core genes than that of singletons shared between the genomes whereas when the circle plot was made between *Geobacillus, Lactobacillus*, and *Bifidobacterium* genomes, they shared very few core genes as shown in **Fig.9; Fig10, Supplementary Fig.2; Fig3**. Phylogenetic Pangenome Accumulation was run to view the pangenome in a phylogenetic context and to determine the entry and exit of gene families in the branch of interest. The Phylogenetic Pangenome Accumulation tree shows the bifurcation of two main clades as shown in the phylogenetic trees discussed above. Node zero was important as it bifurcates into two main clades with node 1 (gives rise to genus *Bifidobacterium* clade) and node 6. Node 6 was also important as it bifurcates into two separate branches of the genus *Geobacillus* as one clade and the genus *Lactobacillus* as another clade. At node zero, there was the maximum number of total genes present (17392) with a very less perfect core of only 0.8%. there is the accumulation of genes at node 1 for genus Bifidobacterium with an increase in perfect core of 27.3%. Similar to node zero there was the second maximum number of total genes (14861) (as the rest of the genes got accumulated into genus *Bifidobacterium*) with a very small perfect core of only 1.6%. This total gene numbers gradually accumulated and shows the increase in perfect core as shown at node 18 (with the perfect core of 8.6%) for

386 the genus *Geobacillus* and at node 7 (with a perfect core of 8.9%) for the genus *Lactobacillus.*
387 These genes further get accumulated within sub-branches as shown in **Fig.11**.

388 **Discussion**

389 The species of the *Geobacillus* genus exhibits ecological, physiological, and genetic diversity,
390 merely due to micro-evolutionary traits such as horizontal gene transfer, which remains a
391 strong factor contributing to the evolution of *Bacillus* to *Geobacillus* (26) Thriving in various
392 harsh conditions thermophiles have been exploited in various industrial and biotechnological
393 applications. However, only a few studies suggest the probiotic nature of the species belonging
394 to the genus *Geobacillus*. Here we have presented the vast *in silico* evaluation of the genus
395 *Geobacillus* considering its probiotic nature. The analysis has been done in genus *Geobacillus*
396 keeping already known probiotic strains of genus *Lactobacillus* and *Bifidobacterium* as
397 positive controls. Various characteristics studied so far for claiming any probiotic species have
398 been covered in this study.

399 Taxonomic and phylogenetic studies revealed that the genomes of *Geobacillus* possess high
400 nucleotide identity and thus are correctly represented in the genus *Geobacillus*. Similar
401 phylogenetic results have been shown earlier for genus *Geobacillus* (27). When considering
402 two other genera such as *Lactobacillus* and *Bifidobacterium*, the phylogenetic trees based on
403 COG-based core genes and single-copy genes using single-copy PGFams specifically form
404 separate clades of each genus thus confirm their correct representation within their respective
405 genus. The annotation results using Prokka suggested that the subsystems highly represented
406 were replication, recombination, repair; post-translational modifications, protein metabolism,
407 and chaperones; nucleotide transport and metabolism; signal transduction mechanism; general
408 functions; energy production and conversion; translation, ribosome structure, and biogenesis;
409 carbohydrate metabolism, amino acid metabolism, and transcription. Whereas the
410 unrepresented features were secondary metabolites biosynthesis, transport, catabolism,
411 intracellular trafficking, secretion, chromatin structure and dynamics, cytoskeleton, and RNA
412 processing and modification. Our research demonstrates that the core genome encompasses
413 the majority of the COG categories. However, flexible genome (accessory and singleton) also
414 shows a considerable percentage as shown in pan-genome accumulation results. Consequently,
415 comprehensive comparative analyses unveiled that crucial functional classes and essential

housekeeping genes remained constant within the core genome. Conversely, genes associated with environmental interaction or the synthesis of secondary metabolites notably exhibited higher abundance within the pan-genome. Also, the considerable flexible genome percentage reveals that Horizontal Gene Transfer (HGT) could represent a crucial mechanism contributing to the environmental adaptive characteristics observed in the genus *Geobacillus*.

Similar findings were reported by Bezuidt et al. in 2016, indicating an overabundance of genes related to COG categories including translation, ribosomal structure, coenzyme transport and metabolism, nucleotide transport and metabolism, as well as protein turnover and chaperones within the conserved core. (26). In a study conducted by Wang et al. in 2020, detailed comparative analyses indicated that core genome stability persisted in fundamental functional classes and essential housekeeping genes, while the flexible genome exhibited a higher prevalence of genes associated with environmental interaction or energy metabolism. Furthermore, instances of horizontal gene transfer (HGT) were identified among various *Geobacillus* species and thus the *Geobacillus* evolution appears to be influenced by environmental factors (28).

As in our study, we found that secondary metabolite metabolism is mainly confined to a flexible genome. Thus, to evaluate this metabolite potential is very important for industrial and biotechnological applications. Wang et al, 2020, shows that various genes for starch, arabinose, glucose, mannose, galactose, and xylose metabolism were also observed in the pan-genome, with a high occurrence of ABC-type sugar transporters that aid in sugar uptake. Therefore, the pan-genome analyses in *Geobacillus* reveal it as a good candidature for hemicellulose degradation (28). A similar overrepresentation of the genes was observed in a study conducted by (26) on pan-genome analyses of 29 *Geobacillus* genomes, depicting that such diverse metabolic machinery aids in its diverse biotechnological potential.

There are various measures for the assessment of any species to be probiotic. The FAO and WHO (2002) guidelines, alongside regulations from the Food and Drug Administration and the Ministry of Public Health in Thailand, delineate the criteria for utilizing probiotic microorganisms in food. These standards encompass specific criteria such as accurate identification, evaluation of probiotic characteristics (such as resistance to gastric acid and bile salts, adherence to mucosal surfaces and epithelial cells, and bile salt hydrolase activity), and

a safety assessment encompassing factors like antimicrobial resistance (AMR), toxin production, and hemolytic activity (29, 30). We have evaluated and covered almost all the characteristics of genus *Geobacillus* to assess its probiotic potential concerning genus *Lactobacillus* and *Bifidobacterium*. Genes related to tolerance to various stress conditions, adhesion, colonization, immune response, etc. governing the probiotic potential were determined. Several genes associated with adhesion, such as *srtD*, *clpB*, and *clpC*, were identified, potentially playing a role in binding to cells or mucosal surfaces within the gut. (31, 32). Other genes such as *TadA*, *PilT*, *PilZ*, and *lapA* essential for pili structure, and thus bacterial colonization were found in the genus *Geobacillus* (33). The key feature defining a bacterium as a probiotic strain hinges on its ability to survive, adapt, or resist low-pH environments (34). Studies suggest that probiotic microorganisms harbor particular genes that aid in enduring or resisting adverse conditions (34). *Lactobacillus* strains typically exhibit resilience to acid and bile, allowing these bacteria to endure and adjust to harsh conditions, rendering them promising candidates for probiotic applications (35). Here we have detected various genes for acid and bile resistance in genus *Geobacillus* such as *sodA*, *luxS*, *glmU*, *atpH*, *recA*, etc. The F1F0-ATPase, encoded by the *atp* operon, typically comprises genes—namely, *atpB*, *atpE*, *atpF*, *atpH*, *atpA*, *atpG*, *atpD*, and *atpC*—in most microbes (36). While most of these genes were identified in the *Geobacillus* genome. These *atp* genes play a crucial role in enabling host microorganisms to survive or withstand acidic environments. The main role of the "*atp*" operon is to facilitate proton pumping, transporting protons from the bacterial cytoplasm outward, and contributing to the maintenance of a neutral pH in the bacterial cytosol. (37).

In response to hyperosmotic stress, organisms accumulate osmotically active compounds. These compounds, known as compatible solutes, do not hinder enzymatic processes within the cell. The gathering of these solutes, known as osmoadaptation, works to counter the outward movement of water, thereby maintaining cell turgor (38). Numerous thoroughly characterized compatible solute transporters have been detected in various gram-positive organisms, such as *Bacillus subtilis* and *Lactococcus lactis*. *OpuA*, *OpuB*, and *OpuC* belong to the ATP binding cassette (ABC) superfamily and are closely associated transporters. These transporters have the capability to transport distinct compounds: *OpuA* transports proline betaine and glycine betaine, *OpuB* transports choline, while *OpuC* transports ectoine, crotonobetaine, γ-

477 butryobetaine, carnitine, choline-O-sulfate, choline, proline betaine, and glycine betaine. (39).
478 These osmotic stress-related genes were absent in *Bifidobacterium* and *Lactobacillus* genomes,
479 however, *opuD* and *opuC* were present in some *Geobacillus* genomes, and *opuCA*, *opuCB*, and
480 *opuCC* were only present in *Geobacillus* genomes. Genes related to oxidative stress such as
481 thioredoxin (*trx*), and ferrous iron transporter (*feoB*) were present in almost all the genomes.
482 However, transcriptional regulator (*oxyR*), and NADH (*ndhH*, *ndhB*, and *ndhC*) were present
483 only in *Geobacillus* genomes. These genes are associated with antioxidant and oxidative stress
484 responses in many bacteria (40, 41).

485 The genes responsible for encoding chaperones (*dnaK*, *groEL*, *groES*) play a crucial role in a
486 broad stress response, involving protection, elimination of damaged proteins, and various
487 related functions. Additionally, the analysis revealed *dnaJ* and *grpE*, which have demonstrated
488 a responsive behavior (upregulation) specifically to acidic environments (42). All the genomes
489 examined contained these molecular chaperones. However, *groES* and *groEL* were present in
490 some *Geobacillus* genomes only. The presence of these probiotic-related genes in the genus
491 *Geobacillus* makes them a promising candidate for exploiting them as probiotic strains.
492 However, culture-dependent analysis of these probiotic features must be validated.

493 Analyzing the mobilome offers a deeper comprehension of genome stability, adaptability, and
494 evolution in host probiotics, particularly in evaluating the potential acquisition and transfer of
495 new genes, including those linked to antibiotic resistance. Nevertheless, if a probiotic strain's
496 genome harbors mobile genetic elements (MGEs) along with antibiotic resistance genes, it
497 becomes unfit for use due to the risk of transferring these resistance genes via processes like
498 conjugation or other mechanisms (43, 44). The lack of mobile genetic elements, virulence
499 factors, and minimal putative insertion sequences indicates the stability of the genus
500 Geobacillus, potentially making it favorable for probiotic candidacy. Regarding the anticipated
501 prophages found in *Geobacillus* genomes, only a small number belonging to the Myoviridae
502 and Siphoviridae families were identified. However, most of these were hypothetical and
503 lacked the typical genes responsible for encoding structural proteins, DNA regulation, lysis,
504 and other essential functions as per the Prokaryotic Virus Orthologous Groups (pVOGs)
505 database. Consequently, these could be classified as defective prophages (45). Plasmids can
506 impart new traits to probiotic bacteria, including enhancements in bacterial metabolism,

adherence, and even antibiotic resistance. However, if plasmids contribute antibiotic resistance to probiotic bacteria, this particular attribute is considered undesirable for a probiotic bacterium (46). However, in our study, we did not find any plasmid in the genomes of genus *Geobacillus*. We have also reported the same results previously in different studies (13, 27). The absence of mobile genetic elements, prophages, and plasmids reveals that the genomes of the genus *Geobacillus* are stable and good candidates as probiotics.

Bacteriocins, a diverse array of ribosomally synthesized antimicrobial peptides, provide probiotic bacteria with the advantageous ability to effectively combat other bacterial strains. They assist in the establishment of a producer bacterium within a specific environment, directly hinder the intrusion of competing strains or pathogens, and have the potential to alter the microbiota's composition while influencing the host's immune system (47). Certain studies have documented the production of bacteriocins within the Bifidobacterium genus (48). Similarly, several bacteriocins, including Circularin_A, ComX1, ComX4, Salivaricin_D, Sactipeptides, Pumilarin, and Geobacillin_I_like, were found across different genomes of *Geobacillus*. CRISPR-Cas systems are adaptive immune systems of microbes. These elements play a crucial role for bacteria in managing phage sequences found in the environment, particularly within the intricate ecosystem of the gut where a diverse viral community thrives. Phages can disrupt bacteria, significantly affecting the survival of bacterial populations, thereby impacting various parameters including the production of probiotics, fermentation duration, taste, and other essential factors (49). Given the resilience of phages to pasteurization and the difficulty in eliminating them, the pursuit of probiotics possessing the capacity to shield against phages and other genetic intruders like plasmids becomes an urgent necessity. In a previous study, 77% of 48 analyzed species of *Bifidobacterium* possess CRISPR-Cas systems (50). In our study, considering two prominent probiotic species of *Bifidobacterium* (*B. breve* and *B. longum*), we didn't find any CRISPR-Cas-related genes. However, there were diverse genes related to CRISPR-Cas systems present in the genomes of *Geobacillus* such as *Cas1*, *Cas2*, *Cas3*, *Cas4*, *Cas5h*, *Cas6*, *Cas9*, *Cmr1*, *Cmr3*, *Cmr4*, *Cmr5*, and *Cmr6*. Thus, the presence of bacteriocins and CRISPR-Cas-related genes suggested the promising potential of probiotics among the genus *Geobacillus*.

A pivotal aspect of defining a probiotic strain is its stance on antibiotic resistance. Within the Qualified Presumption of Safety (QPS) criteria, evaluating probiotics involves examining the safety and antimicrobial resistance (51). Ideally, probiotics ought to be susceptible to at least two antibiotics or should lack inherent antimicrobial resistance (52). The genotypic approach involves full genome sequencing, including plasmids, to detect known antibiotic resistance (AR) genes. If these genes are surrounded by mobile elements or encoded within plasmids, it is advised against commercializing the strain. Safety considerations are crucial, especially if resistance involves clinically important antibiotics like vancomycin. If resistance affects less clinically relevant antibiotics, the risk is lower, but thorough evaluation is needed before commercialization (53). In our previous study (13) we found the genus *Geobacillus* is devoid of any antibiotic resistance. The results collide with the findings of Puopolo and co-workers, where *G. stearothermophilus* GF16 isolated from the active volcanic area was susceptible to different antibiotics used (even at the lowest concentrations of 5 µg/mL) (54). Likewise, the complete genome examination of G. thermoleovorans derived from geothermal springs indicates the lack of antimicrobial resistance genes (55). All of these findings of the absence of antibiotic susceptibility in *Geobacillus* in general (13). Similarly, in the present study, we only found two genes *vanY* and *vanT* related to vancomycin resistance. *VanY* acts as a D, D-carboxypeptidase, eliminating the terminal D-Ala from peptidoglycan to incorporate D-Lactate. The resulting D-Ala-D-Lac peptidoglycan subunits exhibit lower binding affinity with vancomycin in comparison to D-Ala-D-Ala. *VanY* serves as an "accessory" component in the Van cascade and isn't an absolute necessity for vancomycin resistance (56). *Bifidobacterium* and *Lactobacillus* genomes were found to possess more genes apart from *vanY* and *vanT* such as *ImrD*, *rpsL*, and *qacJ*. A study has also reported the presence of erm, *ileS*, and *rpoB* genes in *B. breve*. Investigating the presence of virulence factors (including toxins and enzymes that could potentially heighten the microorganism's pathogenicity) in the genome is crucial for identifying potential safety concerns. However, only a few toxin-like proteins and virulence factors are also absent in all the genomes of the genus *Geobacillus*. A few toxin-like proteins detected were *PhoH*, *MazF*, and *PemK* as compared to genus *Bifidobacterium* and *Lactobacillus* which possess an array of toxin-related genomes such as *FitB*, *PemK*, *MazF*, *ParB*, *RelB*, *RelE*, *MraZ*,*YoeB*, *PhoH*, and *HipA*. *PhoH2* proteins, found in various microorganisms across bacteria and archaea, comprise two domains: An N-terminal PIN-

domain connected to a C-terminal *PhoH* domain. This fusion function as both an RNA helicase and ribonuclease. Mainly within mycobacteria, the genome encodes PhoH2 proteins alongside an adjacent gene, *phoAT,* believed to function as an antitoxin (57). The *mazEF* module, found on *Escherichia coli's* chromosome, serves as a stress-induced toxin-antitoxin system that triggers programmed cell death in *E. coli*. Precisely, *mazF* encodes a resilient toxin, whereas *mazE* encodes a comparatively less stable antitoxin (58). In the *Bacillus anthracis* genome, a toxin-antitoxin (TA) module is present, consisting of *pemI* (antitoxin) and *pemK* (toxin). *PemK* functions as a powerful ribonuclease, demonstrating a preference for pyrimidines (specifically C and U), operating as a translational attenuator (59). Thus, in comparison to already known probiotic candidates of genus *Bifidobacterium* and *Lactobacillus,* the genus *Geobacillus* possess very few toxin-like protein coding genes. Thus, it has been proposed that genus *Geobacillus* may be good candidates with good probiotic potential.

Microbial genomes encode a multitude of genes responsible for Carbohydrate-active enzymes (CAZymes), whereas in humans, only 17 relevant ones have been identified (60). This discrepancy reveals that humans lack extensive enzyme machinery to utilize a wide array of complex carbohydrates. As a result, humans depend on a symbiotic co-metabolism with their microbiota to extract energy, particularly from carbohydrates that are otherwise indigestible (61). The dynamic CAZyme profile is influenced by factors like available carbohydrates, non-carbohydrate food sources, delivery methods, and individual lifestyles (62). Differences in host CAZymes or the lack of particular microbial species possessing distinct CAZymes can substantially impact how the host metabolizes various sources of carbohydrates (63), impacting gut microbiota metabolism and potentially affecting host health. Assessing microbial genetic diversity (CAZy-typing) helps predict which carbohydrates the host can metabolize and identifies underrepresented CAZyme families that might require supplementation through methods like microbiota transplantation or probiotics. In our study, we found an abundance of glycoside hydrolases followed by Glycosyl transferases and carbohydrate esterases among the genus *Geobacillus*. Thus, suggesting that they are capable of enzymatic degradation of various carbohydrates and consequently adding to the probiotic potential of these species. Thus, all the studied in silico features favor the probiotic nature of the genus *Geobacillus*. Yet, an assessment of *in vitro* characterization remains essential concerning probiotic attributes.

The unprecedented probiotic potential of the *Geobacillus* genus remains unexplored. With the Pan-Genome analysis's advent, the intricate genetic makeup, the core genome concerning the essential function genes, and genes required to acclimatize in a new niche or secondary metabolite synthesis (conferred by the flexible genome) in a bacterium can be comprehensively disclosed. Analyzing the Pan-genome serves as a crucial method to provide insights into evolution and adaptation, and to explore potential probiotic characteristics in the group. Comparison to the genus *Lactobacillus* and *Bifidobacterium* in the current study revealed that the genus *Geobacillus* possesses an array of genes that aid in combatting stress conditions. Various genes conferring acidic tolerance and osmotic stress tolerance were found, with certain genes aiding in osmotic and oxidative stress tolerance confined to the genome of *Geobacillus* only. Determining the presence of mobile genetic elements (MGEs) remains paramount to instigating a strain with probiotic potential. Notably, *Geobacillus* appears stable, lacking mobile genetic elements (MGEs), antibiotic resistance genes, and mostly defective prophages. There were no observed plasmids, reducing the risk of horizontal gene transfer for antibiotic resistance. The presence of diverse bacteriocins and an abundance of CRISPR-Cas system genes in *Geobacillus* further suggests its potential against harmful elements like bacteriophages. Additionally, the prevalence of carbohydrate-active enzymes (CAZymes), especially glycoside hydrolases, glycosyl transferases, and carbohydrate esterases, hints at the ability of the genus *Geobacillus* to break down complex carbohydrates. This could enhance its potential as a probiotic, particularly in creating symbiotic food products when combined with prebiotics. While this study unveils promising probiotic traits in the genus *Geobacillus*, further in vitro and in vivo investigations are necessary to confirm these findings. Overall, the comparative genomic analysis sheds light on *Geobacillus*' probiotic potential, yet empirical studies are imperative for substantiating these observations.

Materials and Methods

Genome Availability and Quality Control

All the 13 validly published *Geobacillus* species with correct names as per List of Prokaryotic names with Standing in Nomenclature (LPSN) and 5 randomly selected genomes (not validly published) of genus *Geobacillus* were selected and extracted in Fasta and GenBank formats from NCBI datasets https://www.ncbi.nlm.nih.gov/datasets/genome/. To access the probiotic

potential of various studied *Geobacillus* species, 5 already known probiotic species such as *Lactobacillus acidophilus* ATCC 9224 [NZ_CP130437], *Lactobacillus delbrueckii* subsp. *lactis* DSM 20072 [CP022988], *Lactiplantibacillus plantarum* DSM 20174 [CP039121], *Bifidobacterium breve* DSM 20213 [BBAO00000000], and *Bifidobacterium longum* subsp. *infantis* ATCC 15697 [CP001095] were taken as control species. After the extraction of a total of 23 genomes, they were further analyzed for quality check, trimming, and alignment. The assessment of genome quality was done involved utilizing QUAST (version 4.4) (64). Additionally, to reconfirm genome completeness and identify any contamination, CheckM v1.0.18 (65) was employed, followed by trimming using the Trimmomatic program (version 0.36) (66). Multiple whole genome sequence alignments were conducted using Bowtie2 v2.3.2 (67). These analyses were performed on the KBase platform, available at https://www.kbase.us/ (68).

Phylogenomic Analysis and Annotation

The study examined the phylogenetic relationships among *Geobacillus* species and their connections with other *Lactobacillus* and *Bifidobacterium* probiotic genomes used in this research, based on a defined set of 49 core, universally recognized genes from COG (Clusters of Orthologous Groups) gene families. This was performed by a KBase app Species Tree Builder v0.1.4 (68). This process begins by choosing a subset of publicly available KBase genomes that exhibit close relations to the genomes we've provided. Subsequently, these genomes undergo multiple sequence alignment (MSA) against 49 COG domains. The resulting alignments are meticulously refined using GBLOCKS to eliminate inadequately aligned sections within the MSA. These refined MSAs are concatenated, and a phylogenetic tree is generated using the maximum likelihood method via FastTree2 version 2.1.11. The phylogenomic analysis was also done by using the Codon Tree Test method of the phylogenetic tree-building service of Bacterial and Viral Bioinformatics Resource Center (BV-BRC) (https://www.bv-brc.org/) (69). The Codon Tree approach uses the RAxML program to analyze aligned proteins and coding DNA from single-copy genes after choosing single-copy BV-BRC PGFams. RaxML's "Rapid" bootstrapping option was used to produce support values over 100 iterations.

The genomes were annotated using Prokka v1.14.5 (70) through the KBase platform. To assess the genes carrying known functions among the Prokka annotated genomes was done by the KBase app "View Function Profile for Genome-v1.4.0" and a heatmap was generated. It summarizes gene functions using canonical gene family assignment from COG, PFAM, TIGRFAM, and the SEED. Before this step, the domains of the Prokka annotated genomes were first annotated using the Domain Annotation app v1.0.10. (68).

Detection of Genes Related to Probiotic Features

Genes involved in the mechanism of modulation of the immune system, vitamin biosynthesis, fatty acid synthesis, resistance to stress conditions (acid, bile, osmotic, and oxidative stress), adhesion, bacterial colonization, and molecular chaperones were detected from Prokka annotated genomes (70). These probiotic-related genes were compared with those of already known probiotic species considered in this study.

Detection of Mobile Elements, Insertion Sequences, Plasmids, Prophages, Bacteriocins, and CRISPR-Cas Systems

The in-silico identification of plasmids among the studied genomes was done with PlasmidFinder 2.0 (71). Phage presence was assessed using Phigaro through the Proksee platform. (72). The mobile elements and insertion sequences were predicted by ICEFinder (73) and VRprofile2 (74). The putative bacteriocins were predicted by using BAGEL4 (75), and the presence of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) and Cas proteins were analyzed with the CRISPRCasfinder v4.2.20 tool (76) using Proksee (72) platform and were also cross-checked in Prokka annotations of genomes using KBase platform.

Detection of Antibiotic Resistance Genes, Virulence Factors, and Toxins

The assessment of antibiotic resistance gene predictions was conducted using ABRIcate v1.0.1 within the Galaxy platform, RGI (Resistance Gene Identifier) 6.0.3, and CARD (Comprehensive Antibiotic Resistance Database) (77) using PATRIC (BV-BRC v3.32.13a) and Proksee platforms. The virulence factors were predicted with PATRIC (BV-BRC v3.32.13a) and ABRIcate v1.0.1 against VFDB (virulence factor data base) (69). The toxins were predicted with TAser (https://shiny.bioinformatics.unibe.ch/apps/taser/) using the

TASmania database (78). Only those genes were considered with hmm_E_value less than 1e-15.

Identification of Carbohydrate-active Enzymes (CAZyme)

The carbohydrate-active enzymes were identified by annotating the genome sequences through dbCAN (79) with dbCAN2 HMMs of CAZy families - v10 app in KBase. This method scans protein sequences found in Genomes using a set of Hidden Markov Models (HMMs) from the dbCAN2 CAZy family collection or CAZy database (http://www.cazy.org/). It uses HMMER software v3.3.2 is installed from http://hmmer.org.

Pangenome Analysis

The Pangenome analysis was performed with Genome Comparison SDK v.0.0.7 app in KBase. The KBase algorithm for computing the protein pangenome relies on a k-mer-based approach, providing significant advantages. It entails identifying uniquely distinct k-mers, specifically of length 8, present in each protein across all genomes. Phylogenetic Pangenome Accumulation v1.4.0 (https://kbase.us/applist/apps/kb_phylogenomics/view_pan_phylo/) was run to view the pangenome in phylogenetic context and to compare among two genome sets (between *Geobacillus* and *Lactobacillus/Bifidobacterium*). This application allows to dissecting of pangenome categories using a species tree to determine the entry and exit of gene families in the branch of interest.

The Pangenome analysis was also done with IPGA (Integrated Prokaryotes Genome and Pan-Genome Analysis) v1.09 (80). Before any analysis quality control is done on input genomes. After QC, the genes of all filtered genomes are predicted using prokka. Using up to eight different types of software such as OrthoMCL, PanOCT, Roary, OrthoFinder, panX, Panaroo, PPanGGoLiN, and PEPPAN each gene is annotated against the COG database to provide a unique pan-genome profile. To assist users in choosing the optimal pan-genome profile from the possibly inconsistent findings, IPGA extracts all orthologous gene clusters and then assigns a score to each of them (80). Pangenome Circle Plot was formed using Pangenome Circle Plot-v1.2.0 in KBase (https://kbase.us/applist/apps/kb_phylogenomics/view_pan_circle_plot/release). This app allows the overlapping membership of genes against a base genome.

Acknowledgments: Dr. Ishfaq Nabi Najar would like to thank DBT-RA, Department of Biotechnology, Govt. of India, for providing the DBT-RA Fellowship (**DBT-RA/2022/JULY/N/2877**) for research work. I would also like to thank CSIR IIIM Jammu, for providing me lab space to carry out my research.

Ethics approval: Not applicable

Funding: Not applicable.

Credit authorship contribution statement:

Ishfaq Nabi Najar: Conceptualization, Data Curation, Analysis, Writing – original draft. **Prayatna Sharma:** Analysis and Writing– original draft. **Rohit Das:** Analysis and Editing – original draft. **Krishnendu Mondal, Ashish Kumar Singh**, **Anu Radha**, **Varsha Sharma**, **Sonali Sharma-** Data Curation, Writing and Editing – original draft, **Nagendra Thakur**: Review & Editing. **Sumit G Gandhi**: Review & Editing. **Vinod Kumar:** Conceptualization, Review & Editing.

Declaration of Competing Interest: The authors declare no competing interests.

References:

Najar IN, Thakur N. 2020. A systematic review of the genera *Geobacillus* and *Parageobacillus*: Their evolution, current taxonomic status, and major applications. Microbiology 166: 800–816. https://doi.org/10.1099/mic.0.000945.

Khaswal A, Chaturvedi N, Mishra SK, Kumar PR, Paul PK. 2022. Current status and applications of genus *Geobacillus* in the production of industrially important products—a review. Folia Microbiol 67: 389–404. https://doi.org/10.1007/s12223-022-00961-w.

Novik G, Savich V, Meerovskaya O. 2019. *Geobacillus* Bacteria: Potential Commercial Applications in Industry, Bioremediation, and Bioenergy Production Growing and Handling of Bacterial Cultures. IntechOpen.

Cortés-Antiquera R, Márquez SL, Espina G, Sánchez-SanMartín J, Blamey JM. 2023. Recombinant expression and characterization of a new laccase, bioinformatically identified, from the Antarctic thermophilic bacterium *Geobacillus* sp. ID17. Extremophiles 27.

743 Meslé MM, Mueller RC, Peach J, Eilers B, Tripet BP, Bothner B, Copié V, Peyton BM. 2022.
744 Isolation and Characterization of Lignocellulose-Degrading *Geobacillus thermoleovorans* from
745 Yellowstone National Park. Appl Environ Microbiol 88.

746 Wada K, Suzuki H. 2020. Biotechnological platforms of the moderate thermophiles, *Geobacillus*
747 species: notable properties and genetic tools, p. 195–218. *In* Physiological and Biotechnological
748 Aspects of Extremophiles. Elsevier.

749 Ash C, Farrow JAE, Wallbanks S, Collins MD. 1991. Phylogenetic heterogeneity of the genus
750 *Bacillus* revealed by comparative analysis of small-subunit-ribosomal RNA sequences. Lett Appl
751 Microbiol 13:202–206.

752 Lin JH, Zhang KC, Tao WY, Wang D, Li S. 2019. *Geobacillus* strains that have potential value in
753 microbial-enhanced oil recovery. Appl Microbiol Biotechnol 103: 8339–8350.
754 https://doi.org/10.1007/s00253-019-10115-7.

755 Donk PJ. 1920. A highly resistant thermophilic organism. J Bacteriol 5:373–374.

756 Nazina TN, Tourova TP, Poltaraus AB, Novikova E V, Grigoryan AA, Ivanova AE, Lysenko AM,
757 Petrunyaka V V, Osipov GA, Belyaev SS, MV I. 2001. Taxonomic study of aerobic thermophilic
758 bacilli: descriptions of *Geobacillus subterraneus* gen. nov., sp. nov. and *Geobacillus uzenensis* sp.
759 nov. from petroleum reservoirs and transfer of *Bacillus stearothermophilus*, *B. thermocatenulatus*,
760 *B. thermoleovorans*. Int J Syst Evol Microbiol 51:433–446.

761 Logan NA. 2014. The genus *Geobacillus*, p. 133–147. *In* The Prokaryotes: Firmicutes and
762 Tenericutes. Springer-Verlag Berlin Heidelberg.

763 Zebrowska J, Witkowska M, Struck A, Laszuk PE, Raczuk E, Ponikowska M, Skowron PM,
764 Zylicz-Stachula A. 2022. Antimicrobial Potential of the Genera *Geobacillus* and *Parageobacillus*,
765 as Well as Endolysins Biosynthesized by Their Bacteriophages. Antibiotics 11.

766 Najar IN, Das S, Kumar S, Sharma P. 2022. Coexistence of Heavy Metal Tolerance and Antibiotic
767 Resistance in Thermophilic Bacteria Belonging to Genus *Geobacillus*. Front Microbiol 13:1–15.

768 Jezewska-Frackowiak J, Seroczynska K, Banaszczyk J, Jedrzejczak G, Zylicz-Stachula A,
769 Skowron PM. 2018. The promises and risks of probiotic *Bacillus* species. Acta Biochim Pol
770 65:509–519.

771 Łubkowska B, Jeżewska-frąckowiak J, Sobolewski I, Skowron PM. 2021. Bacteriophages of
772 thermophilic '*Bacillus* group' bacteria—a review. Microorganisms 9: 1-27.
773 https://doi.org/10.3390/microorganisms9071522.

774 Mahdhi A, Hmila Z, Behi A, Bakhrouf A. 2011. Preliminary characterization of the probiotic
775 properties of *Candida famata* and *Geobacillus thermoleovorans*. Iranian journal of microbiology,
776 3(3), 129–134.

777 Mcmullan G, Christie JM, Rahman TJ, Banat IM, Ternan NG, Marchant R. 2004. Habitat,
778 applications and genomics of the aerobic, thermophilic genus *Geobacillus*. Biochemical Society
779 transactions, 32: 214–217. https://doi.org/10.1042/bst0320214.

780 Najar IN, Sherpa MT, Das S, Das S, Thakur N. 2020. Diversity analysis and metagenomic insights
781 into antibiotic and metal resistance among Himalayan hot spring bacteriobiome insinuating
782 inherent environmental baseline levels of antibiotic and metal tolerance. J Glob Antimicrob Resist
783 21:342–352.

784 Donkor ES. 2013. Sequencing of Bacterial Genomes: Principles and Insights into Pathogenesis
785 and Development of Antibiotics. Genes (Basel) 4:556–572.

786 Saraswathy N, Ramalingam P. 2011. Genome sequencing methods. Concepts and Techniques in
787 Genomics and Proteomics, 95–107. doi:10.1533/9781908818058.95, p. 95–107. *In* Saraswathy,
788 N, Ramalingam, P (eds.), Concepts and Techniques in Genomics and Proteomics. Woodhead
789 Publishing.

790 Najar IN, Sherpa MT, Das S, Thakur N. 2021. The draft genome sequence of *Parageobacillus* sp.
791 strain SY1 gives insights into its physiological properties and protease production. Meta Gene
792 29:100894.

793 Najar IN, Sherpa MT, Das S, Thakur N. 2020. Bacterial diversity and functional metagenomics
794 expounding the diversity of xenobiotics, stress, defense and CRISPR gene ontology providing eco-
795 efficiency to Himalayan Hot Springs. Funct Integr Genomics 20:479–496.

796 Caputo A, Fournier PE, Raoult D. 2019. Genome and pan-genome analysis to classify emerging
797 bacteria. Biol Direct 14. https://doi.org/10.1186/s13062-019-0234-0.

798 Tiwary BK. 2020. Chapter 3 - Evolutionary pan-genomics and applications, p. 65–80. *In* Barh, D,
799 Soares, S, Tiwari, S, Azevedo, V (eds.), Pan-genomics: Applications, Challenges, and Future
800 Prospects. Academic Press.

801 Asif M, Li-Qun Z, Zeng Q, Atiq M, Ahmad K, Tariq A, Al-Ansari N, Blom J, Fenske L, Alodaini
802 HA, Hatamleh AA. 2023. Comprehensive genomic analysis of *Bacillus paralicheniformis* strain
803 BP9, pan-genomic and genetic basis of biocontrol mechanism. Comput Struct Biotechnol J
804 21:4647–4662.

805 Bezuidt OK, Pierneef R, Gomri AM, Adesioye F, Makhalanyane TP, Kharroub K, Cowan DA.
806 2016. The *Geobacillus* pan-genome: Implications for the evolution of the genus. Front Microbiol
807 7:1–9.

808 Najar IN, Das S, Thakur N. 2020. Reclassification of *Geobacillus galactosidasius* and Geobacillus
809 yumthangensis as *Parageobacillus galactosidasius* comb. nov. and *Parageobacillus*
810 *yumthangensis* comb. nov., respectively. Int J Syst Evol Microbiol 70:6518–6523.

811 Wang M, Zhu H, Kong Z, Li T, Ma L, Liu D, Shen Q. 2020. Pan-genome analyses of *Geobacillus*
812 spp. reveal genetic characteristics and composting potential. Int J Mol Sci 21:1–16.

813 Foongsawat N, Sunthornthummas S, Nantavisai K, Surachat K, Rangsiruji A, Sarawaneeyaruk S,
814 Insian K, Sukontasing S, Suwannasai N, Pringsulaka O. 2023. Isolation, Characterization, and
815 Comparative Genomics of the Novel Potential Probiotics from Canine Feces. Food Sci Anim
816 Resour 43:685–702.

817. Binda S, Hill C, Johansen E, Obis D, Pot B, Sanders ME, Tremblay A, Ouwehand AC. 2020. Criteria to Qualify Microorganisms as "Probiotic" in Foods and Dietary Supplements. Front Microbiol 11:1–9.

820. Zhao X, Wang Y, Shang Q, Li Y, Hao H, Zhang Y, Guo Z, Yang G, Xie Z, Wang R. 2015. Collagen-like proteins (ClpA, ClpB, ClpC, and ClpD) are required for biofilm formation and adhesion to plant roots by *Bacillus amyloliquefaciens* FZB42. PLoS One 10:1–16.

823. Lee J, Sullivan DJO. 2010. Genomic Insights into *Bifidobacteria*. Microbiology and Molecular Biology Reviews 74:378–416.

825. Connell MO, Zomer A, Leahy SC, Reunanen J, Bottacini F. 2011. Functional genome analysis of *Bifidobacterium breve* UCC2003 reveals type IVb tight adherence (Tad) pili as an essential and conserved host-colonization factor. PNAS 108:11217–11222.

828. Nguyen TL, Kim D. 2018. Genome-Wide Comparison Reveals a Probiotic Strain *Lactococcus lactis* WFLU12 Isolated from the Gastrointestinal Tract of Olive Flounder (*Paralichthys olivaceus*) Harboring Genes supporting probiotic action. Mar Drugs 16:1–17.

831. Chou L, Weimer B. 1999. Isolation and Characterization of Acid- and Bile-Tolerant Isolates from Strains of *Lactobacillus acidophilus*. J Dairy Sci 82:23–31.

833. Ventura M, Canchaya C, Sinderen D Van, Fitzgerald GF, Zink R. 2004. *Bifidobacterium lactis* DSM 10140: Identification of the atp (atpBEFHAGDC) Operon and Analysis of Its Genetic Structure, Characteristics, and Phylogeny. Appl Environ Microbiol 70:3110–3121.

836. Duary RK, Batish VK, Grover S. 2010. Expression of the atpD gene in probiotic *Lactobacillus plantarum* strains under in vitro acidic conditions using RT-qPCR. Res Microbiol 161:399–405.

838. Fraser KR, Harvie D, Coote PJ, O'Byrne CP. 2000. Identification and characterization of an ATP binding cassette L-carnitine transporter in *Listeria monocytogenes*. Appl Environ Microbiol 66:4696–4704.

841. Kempf B, Bremer E. 1995. OpuA, an osmotically regulated binding protein-dependent transport system for the osmoprotectant glycine betaine in *Bacillus subtilis*. Journal of Biological Chemistry 270:16701–16713.

844. Shamsuzzaman M, Dahal RH, Kim S, Kim J. 2023. Genome insight and probiotic potential of three novel species of the genus *Corynebacterium*. Front Microbiol 14:1–14.

846. Méndez V, Rodríguez-Castro L, Durán RE, Padrón G, Seeger M. 2022. The OxyR and SoxR transcriptional regulators are involved in a broad oxidative stress response in *Paraburkholderia xenovorans* LB400. Biol Res 55:1–19.

849. Pfeiler EA, Azcarate-Peril MA, Klaenhammer TR. 2007. Characterization of a novel bile-inducible operon encoding a two-component regulatory system in *Lactobacillus acidophilus*. J Bacteriol 189:4624–4634.

Bačun-Družina V, Mrvčić J, Butorac AS, Stehlik-Tomas V, Gjuračić K. 2009. The influence of gene transfer on the lactic acid bacteria evolution. Mljekarstvo 59:181–192.

Sylvere N, Mustopa AZ, Budiarti S, Meilina L, Hertati A, Handayani I. 2023. Whole-genome sequence analysis and probiotic characteristics of *Lactococcus lactis* Subsp. lactis strain Lac3 isolated from traditional fermented buffalo milk (Dadih). Journal of Genetic Engineering and Biotechnology 21:1–12.

Valdez-Baez J, Morais F, Da Costa R, Pinto Gomide AC, Profeta R, Lima Da Silva A, De T, Sousa J, Vinícius M, Viana C, Kato RB, Ferrary Americo M, Dos A, Freitas S, Dias De Oliveira Carvalho R, Brenig B, Martins S, Aburjaile F, Azevedo V. 2022. Comparative Genomics and In Silico Evaluation of Genes Related to the Probiotic Potential of Bifidobacterium breve 110 1A. Bacteria 2022:161–182.

Abriouel H, Pérez Montoro B, de la Fuente Ordoñez JJ, Lavilla Lerma L, Knapp CW, Benomar N. 2019. New insights into the role of plasmids from probiotic *Lactobacillus pentosus* MP-10 in Aloreña table olive brine fermentation. Sci Rep 9:1–19.

Dobson A, Cotter PD, Paul Ross R, Hill C. 2012. Bacteriocin production: A probiotic trait? Appl Environ Microbiol 78:1–6.

Sarkar A, Mandal S. 2016. *Bifidobacteria*—Insight into clinical outcomes and mechanisms of its probiotic action. Microbiol Res 192:159–171.

Yang L, Li W, Ujiroghene OJ, Yang Y, Lu J, Zhang S, Pang X, Lv J. 2020. Occurrence and Diversity of CRISPR Loci in *Lactobacillus casei* Group. Front Microbiol 11:1–9.

Briner AE, Lugli GA, Milani C, Duranti S, Turroni F, Gueimonde M, Margolles A, Van Sinderen D, Ventura M, Barrangou R. 2015. Occurrence and diversity of CRISPR-Cas systems in the genus *Bifidobacterium*. PLoS One 10:1–16.

Koutsoumanis K, Allende A, Alvarez-Ordóñez A, Bolton D, Bover-Cid S, Chemaly M, Davies R, De Cesare A, Hilbert F, Lindqvist R, Nauta M, Peixe L, Ru G, Simmons M, Skandamis P, Suffredini E, Cocconcelli PS, Fernández Escámez PS, Prieto-Maradona M, Querol A, Sijtsma L, Evaristo Suarez J, Sundh I, Vlak J, Barizzone F, Hempen M, Herman L. 2022. Update of the list of QPS-recommended biological agents intentionally added to food or feed as notified to EFSA 15: suitability of taxonomic units notified to EFSA until September 2021. EFSA Journal 20:1–40.

Borriello SP, Hammes WP, Holzapfel W, Marteau P, Schrezenmeir J, Vaara M, Valtonen V. 2003. Safety of probiotics that contain *Lactobacilli* or *Bifidobacteria*. Clinical Infectious Diseases 36:775–780.

Merenstein D, Pot B, Leyer G, Ouwehand AC, Preidis GA, Elkins CA, Hill C, Lewis ZT, Shane AL, Zmora N, Petrova MI, Collado MC, Morelli L, Montoya GA, Szajewska H, Tancredi DJ, Sanders ME. 2023. Emerging issues in probiotic safety: 2023 perspectives. Gut Microbes 15:1–22.

Puopolo R, Gallo G, Mormone A, Limauro D, Contursi P, Piochi M, Bartolucci S, Fiorentino G. 2020. Identification of a new heavy-metal-resistant strain of *Geobacillus stearothermophilus* isolated from a hydrothermally active volcanic area in southern Italy. Int J Environ Res Public Health 17.

Oztug M, Cebeci A, Mumcu H, Akgoz M, Karaguler NG. 2020. Whole-Genome Sequence of *Geobacillus thermoleovorans* ARTRW1, Isolated from Armutlu Geothermal Spring, Turkey Microbiol Resour Announc 9: e00269-20. https://doi.org/10.1128/MRA .00269-20.

Arthur M, Depardieu F, Snaith HA, Reynolds PE, Courvalin P. 1994. Contribution of vanY D, D-carboxypeptidase to glycopeptide resistance in *Enterococcus faecalis* by hydrolysis of peptidoglycan precursors. Antimicrob Agents Chemother 38:1899–1903.

Andrews ESV, Arcus VL. 2020. PhoH2 proteins couple RNA helicase and RNAse activities. Protein Science 29:883–892.

Amitai S, Yassin Y, Engelberg-Kulka H. 2004. MazF-mediated cell death in *Escherichia coli*: A point of no return. J Bacteriol 186:8295–8300.

Agarwal S, Mishra NK, Bhatnagar S, Bhatnagar R. 2010. PemK toxin of *Bacillus anthracis* is a ribonuclease: An insight into its active site, structure, and function. Journal of Biological Chemistry 285:7254–7270.

Kaoutari A El, Armougom F, Gordon JI, Raoult D, Henrissat B. 2013. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. Nat Rev Microbiol 11:497–504.

Soverini M, Turroni S, Biagi E, Quercia S, Brigidi P, Candela M, Rampelli S. 2017. Variation of carbohydrate-active enzyme patterns in the gut microbiota of Italian healthy subjects and type 2 diabetes patients. Front Microbiol 8:1–8.

Ye L, Das P, Li P, Ji B, Nielsen J. 2019. Carbohydrate active enzymes are affected by diet transition from milk to solid food in infant gut microbiota. FEMS Microbiol Ecol 95:1–9.

Aakko J, Pietilä S, Toivonen R, Rokka A, Mokkala K, Laitinen K, Elo L, Hänninen A. 2020. A carbohydrate-active enzyme (CAZy) profile links the successful metabolic specialization of *Prevotella* to its abundance in gut microbiota. Sci Rep 10:1–12.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: Quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM : assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.

68. Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, Dehal P, Ware D, Perez F, Canon S, Sneddon MW, Henderson ML, Riehl WJ, Murphy-Olson D, Chan SY, Kamimura RT, Kumari S, Drake MM, Brettin TS, Glass EM, Chivian D, Gunter D, Weston DJ, Allen BH, Baumohl J, Best AA, Bowen B, Brenner SE, Bun CC, Chandonia JM, Chia JM, Colasanti R, Conrad N, Davis JJ, Davison BH, Dejongh M, Devoid S, Dietrich E, Dubchak I, Edirisinghe JN, Fang G, Faria JP, Frybarger PM, Gerlach W, Gerstein M, Greiner A, Gurtowski J, Haun HL, He F, Jain R, Joachimiak MP, Keegan KP, Kondo S, Kumar V, Land ML, Meyer F, Mills M, Novichkov PS, Oh T, Olsen GJ, Olson R, Parrello B, Pasternak S, Pearson E, Poon SS, Price GA, Ramakrishnan S, Ranjan P, Ronald PC, Schatz MC, Seaver SMD, Shukla M, Sutormin RA, Syed MH, Thomason J, Tintle NL, Wang D, Xia F, Yoo H, Yoo S, Yu D. 2018. KBase: The United States Department of Energy Systems biology knowledgebase. Nat Biotechnol 36:566–569.

69. Olson RD, Assaf R, Brettin T, Conrad N, Cucinell C, Davis JJ, Dempsey DM, Dickerman A, Dietrich EM, Kenyon RW, Kuscuoglu M, Lefkowitz EJ, Lu J, Machi D, Macken C, Mao C, Niewiadomska A, Nguyen M, Olsen GJ, Overbeek JC, Parrello B, Parrello V, Porter JS, Pusch GD, Shukla M, Singh I, Stewart L, Tan G, Thomas C, VanOeffelen M, Vonstein V, Wallace ZS, Warren AS, Wattam AR, Xia F, Yoo H, Zhang Y, Zmasek CM, Scheuermann RH, Stevens RL. 2023. Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD, and ViPR. Nucleic Acids Res 51: D678–D689.

70. Seemann T. 2014. Prokka: Rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069.

71. Carattoli A, Zankari E, Garcia-Fernandez A, Larsen MV, Lund O, Villa L, Aarestrup FM, Hasman H. 2014. In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother 58:3895–3903.

72. Grant JR, Enns E, Marinier E, Mandal A, Herman EK, Chen C, Graham M, Domselaar G Van, Stothard P. 2023. Proksee: in-depth characterization and visualization of bacterial genomes. Nucleic Acids Res 51:484–492.

73. Liu M, Li X, Xie Y, Bi D, Sun J, Li J, Tai C, Deng Z, Ou HY. 2019. ICEberg 2.0: An updated database of bacterial integrative and conjugative elements. Nucleic Acids Res 47: D660–D665.

74. Wang M, Goh YX, Tai C, Wang H, Deng Z, Ou HY. 2022. VRprofile2: Detection of antibiotic resistance-associated mobilome in bacterial pathogens. Nucleic Acids Res 50: W768–W773.

75. Van Heel AJ, De Jong A, Song C, Viel JH, Kok J, Kuipers OP. 2018. BAGEL4: A user-friendly web server to thoroughly mine RiPPs and bacteriocins. Nucleic Acids Res 46: W278–W281.

76. Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, Néron B, Rocha EPC, Vergnaud G, Gautheret D, Pourcel C. 2018. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance, and integrates search for Cas proteins. Nucleic Acids Res 46: W246–W251.

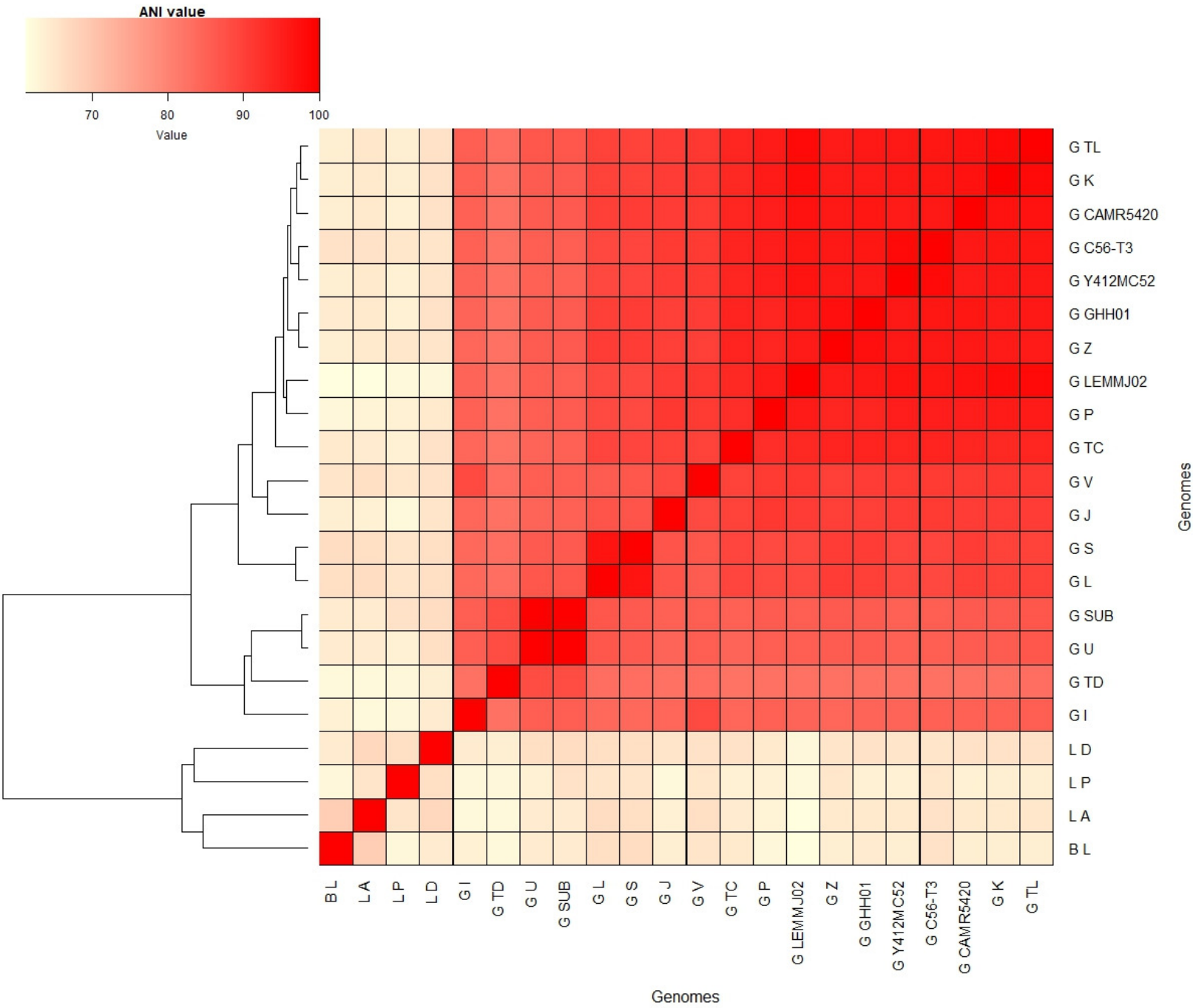
77. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, Huynh W, Nguyen AL V., Cheng AA, Liu S, Min SY, Miroshnichenko A, Tran HK, Werfalli RE, Nasir JA, Oloni M, Speicher DJ, Florescu A, Singh B, Faltyn M, Hernandez-Koutoucheva A, Sharma AN, Bordeleau

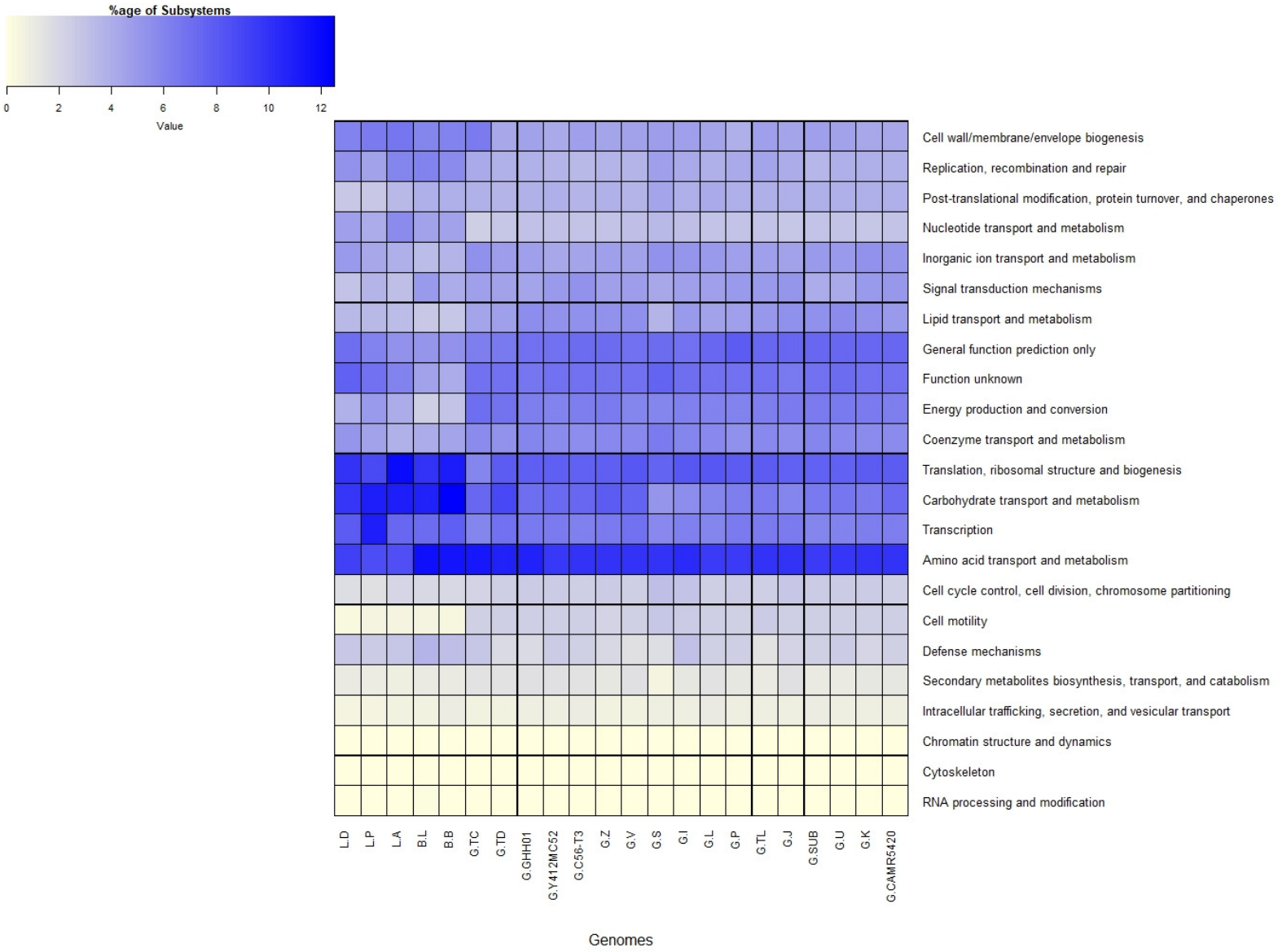
E, Pawlowski AC, Zubyk HL, Dooley D, Griffiths E, Maguire F, Winsor GL, Beiko RG, Brinkman FSL, Hsiao WWL, Domselaar G V., McArthur AG. 2020. CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res 48: D517–D525.

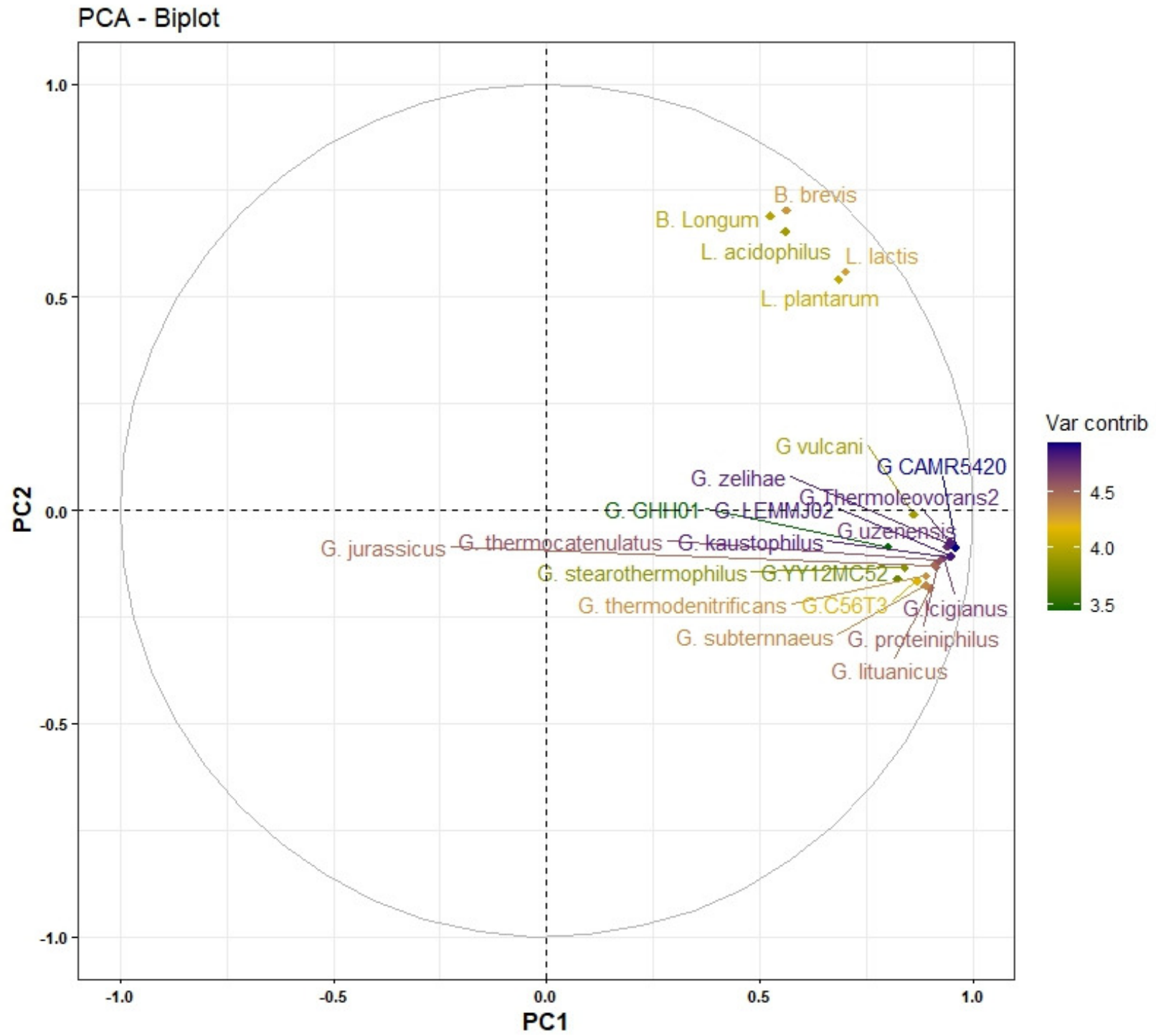
Akarsu H, Bordes P, Mansour M, Bigot DJ, Genevaux P, Falquet L. 2019. TASmania: A bacterial toxin-antitoxin systems database. PLoS Comput Biol 15:1–28.

Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. 2012. DbCAN: A web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 40:445–451.

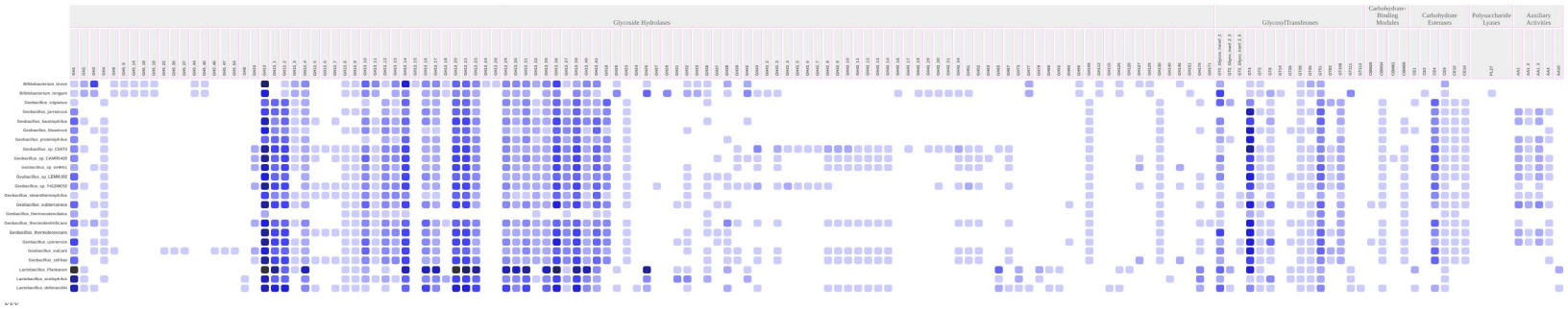
Liu D, Zhang Y, Fan G, Sun D, Zhang X, Yu Z, Wang J, Wu L, Shi W, Ma J. 2022. IPGA: A handy integrated prokaryotes genome and pan-genome analysis web service. iMeta 1:1–7.

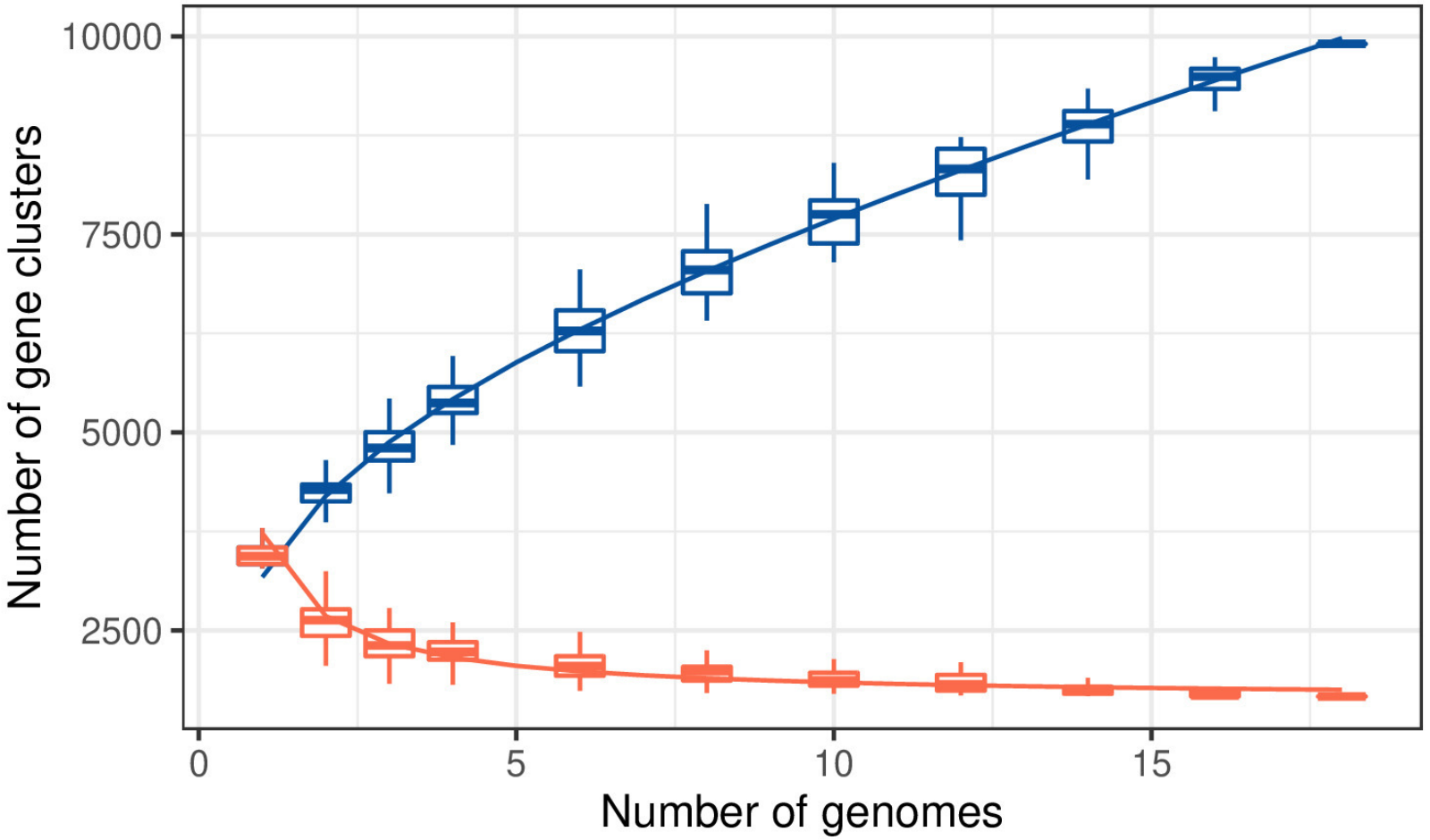


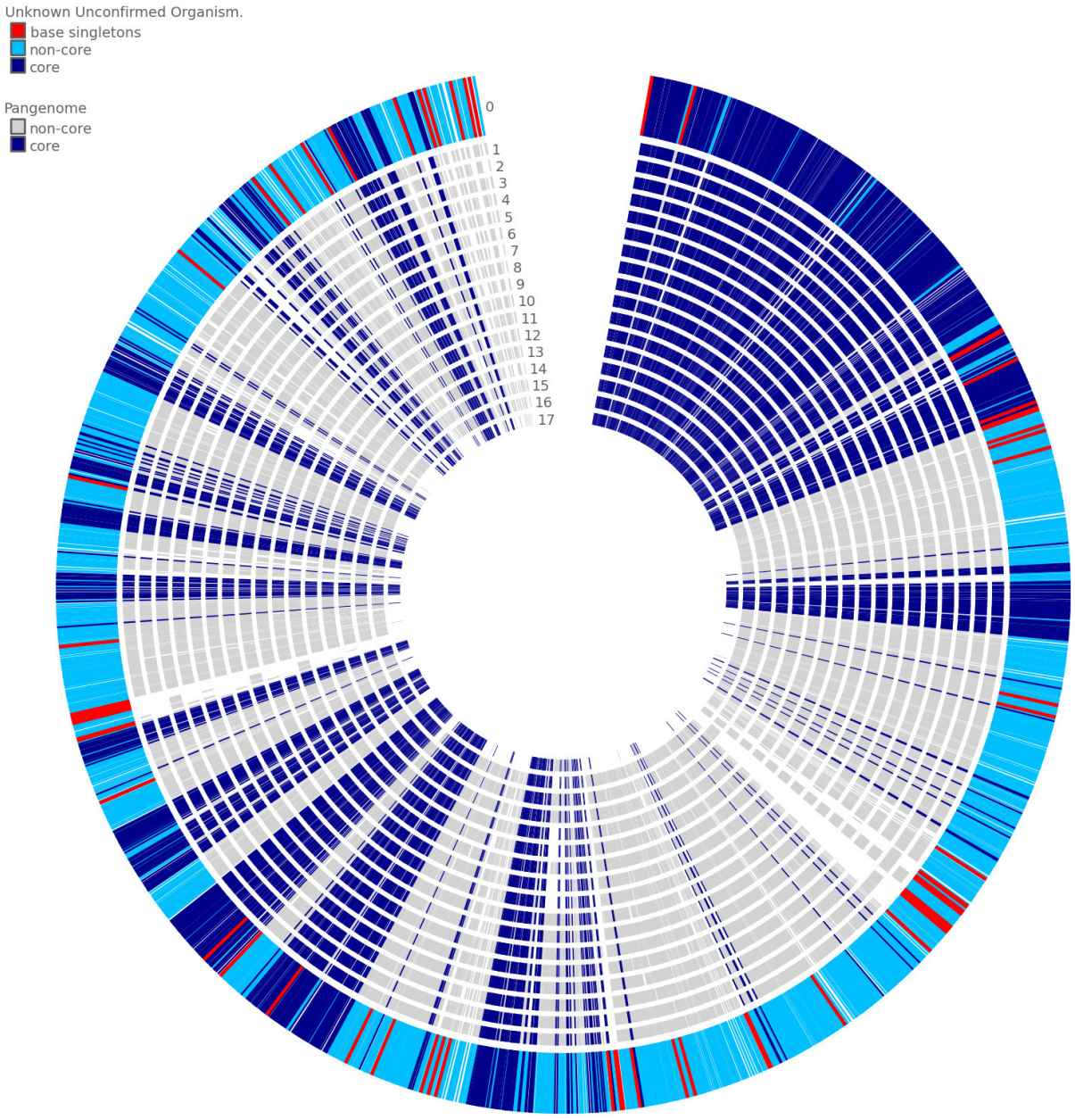


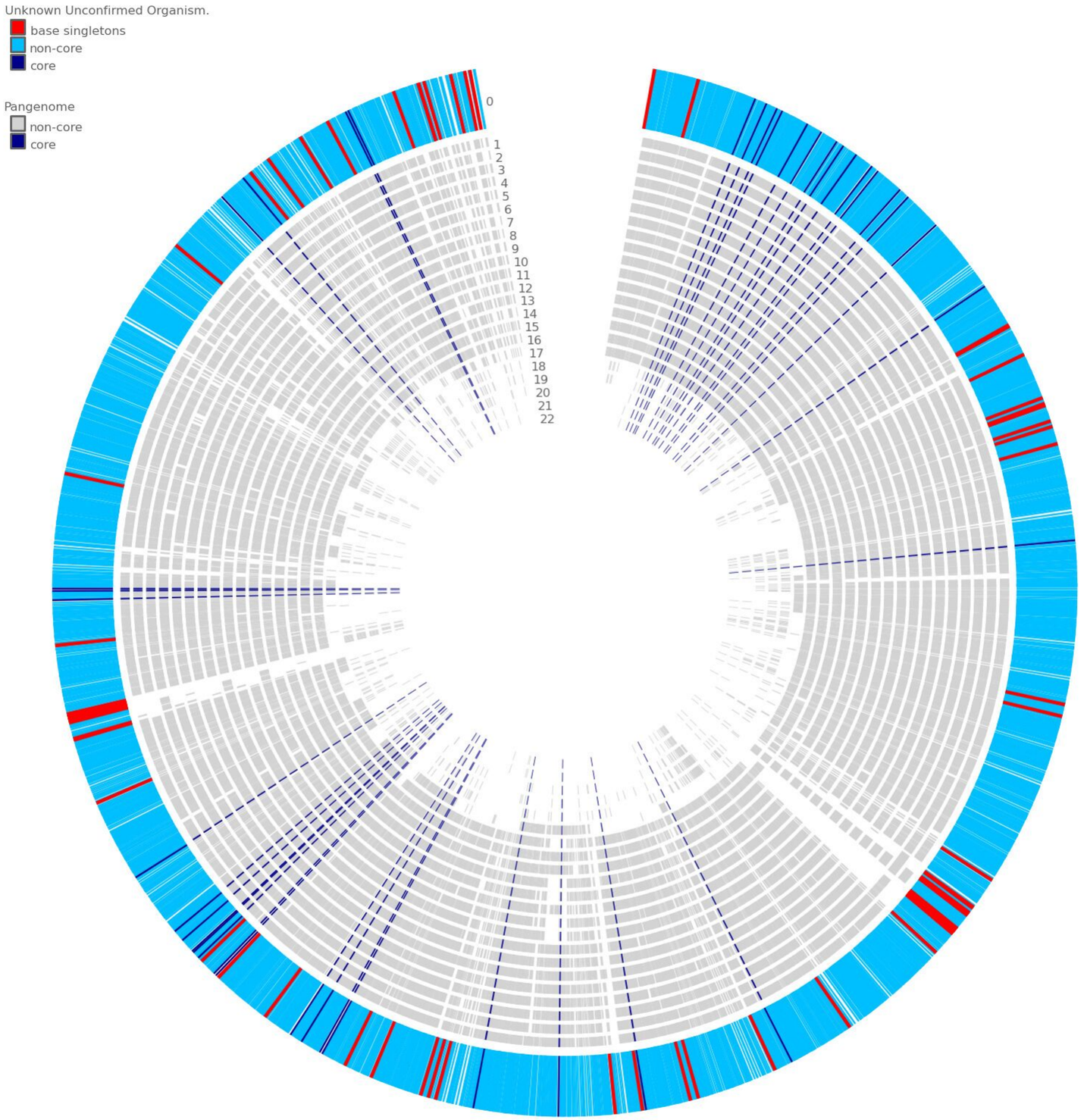


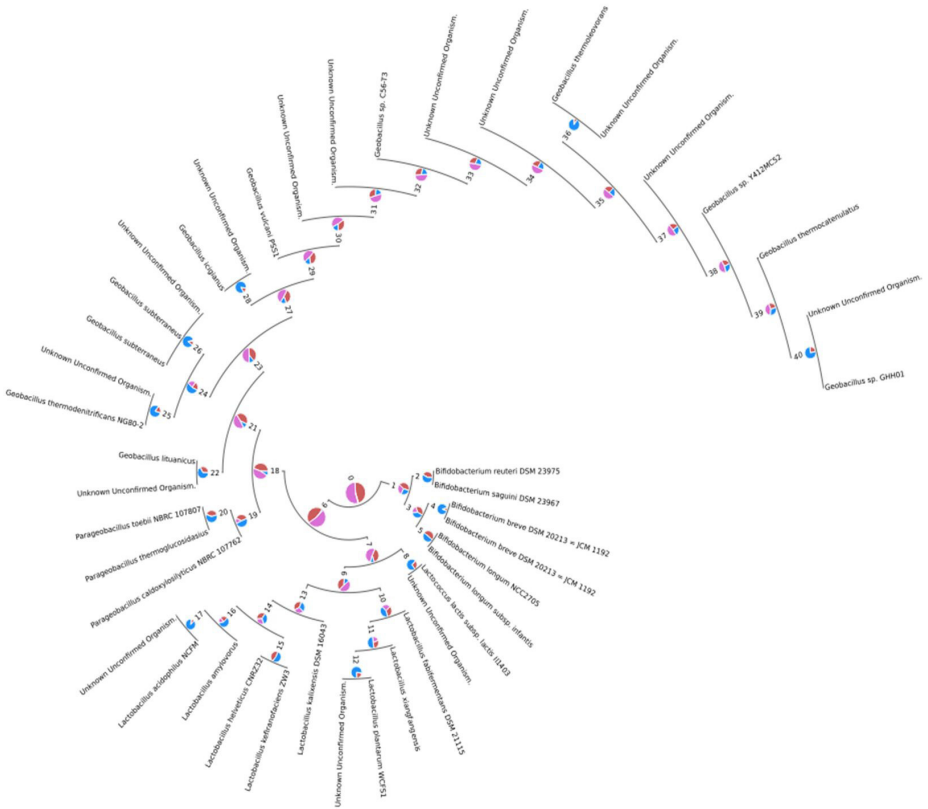
TABULAR PROFILE | Cellulosome HITS | Glycoside Hydrolases HITS | GlycosylTransferases HITS | Carbohydrate-Binding Modules HITS | Carbohydrate Esterases HITS | Polysaccharide Lyases HITS | Auxiliary Activities HITS







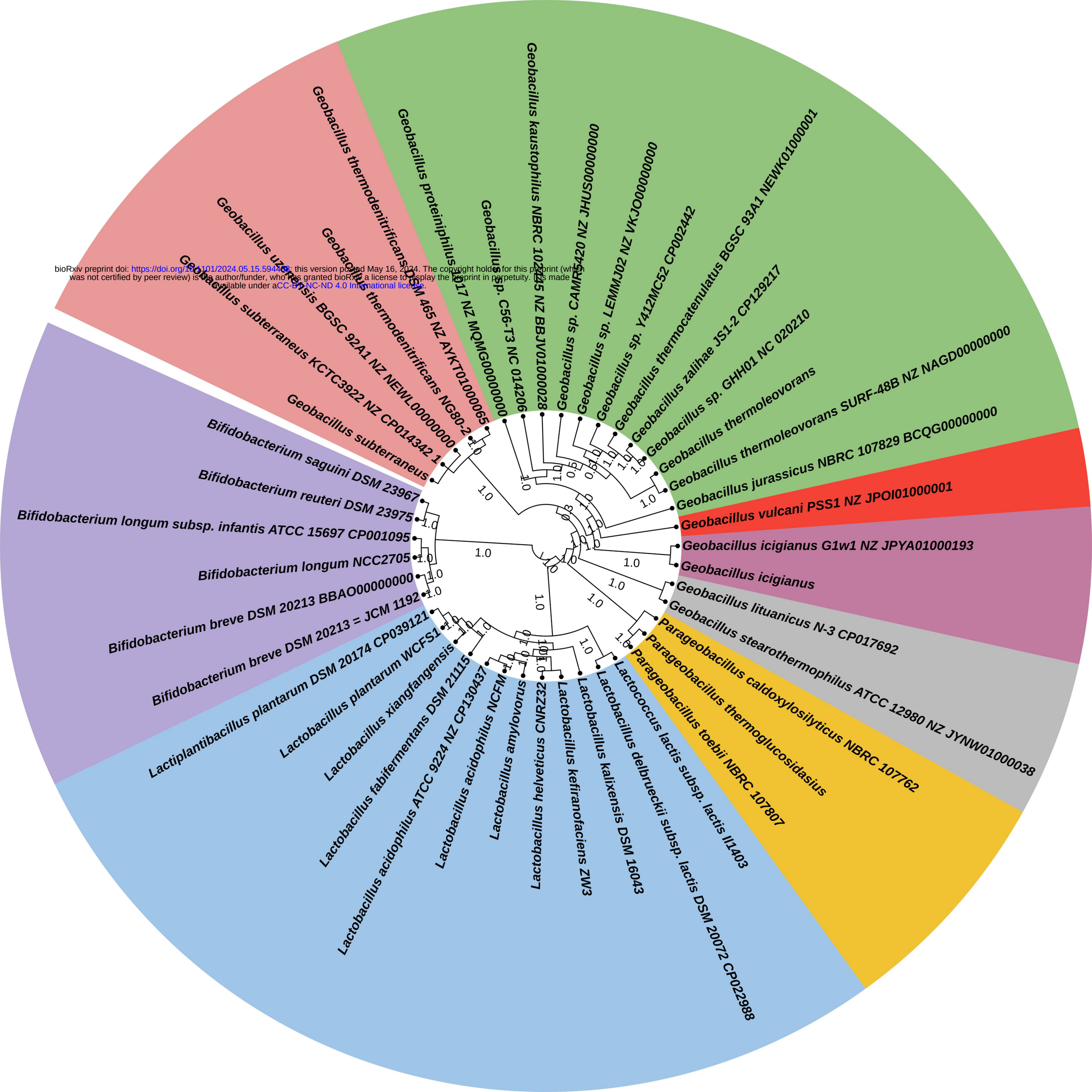

Unknown Unconfirmed Organism.
base singletons
non-core
core
Pangenome
non-core
core
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22



		Total	Singleton	Partial	Perfect Core			Total	Singleton	Partial	Perfect Core
Node 0	Gene No	17392	7974	9279	139	**Node 11**	Gene No	3676	1068	733	1875
	%	100	45.8	53.4	0.8		%	100	29.1	19.9	51
Node 1	Gene No	3945	1687	1180	1078	**Node 12**	Gene No	3204	718	0	2486
	%	100	42.8	29.9	27.3		%	100	22.4	0	77.6
Node 2	Gene No	2567	1108	0	1459	**Node 13**	Gene No	3240	1244	1005	991
	%	100	43.2	0	56.8		%	100	38.4	31	30.6
Node 3	Gene No	3085	1194	645	1246	**Node 14**	Gene No	2953	1050	858	1045
	%	100	38.7	20.9	40.4		%	100	35.6	29	35.4
Node 4	Gene No	1936	194	0	1742	**Node 15**	Gene No	2248	1034	0	1214
	%	100	10	0	90		%	100	46	0	54
Node 5	Gene No	2698	1335	0	1363	**Node 16**	Gene No	2220	513	583	1324
	%	100	49.5	0	50.5		%	100	23.1	17.3	59.6
Node 6	Gene No	14861	6674	7956	231	**Node 17**	Gene No	1844	201	0	1643
	%	100	44.9	53.5	1.6		%	100	10.9	0	89.1
Node 7	Gene No	6998	2733	3641	624	**Node 18**	Gene No	9590	4273	4489	828
	%	100	39.1	52	8.9		%	100	44.6	46.8	8.6
Node 8	Gene No	2458	692	0	1766	**Node 19**	Gene No	4721	1728	695	2298
	%	100	28.2	0	71.8		%	100	36.6	14.7	48.7
Node 9	Gene No	5943	2297	2884	762	**Node 20**	Gene No	4101	1637	0	2464
	%	100	38.7	48.5	12.8		%	100	39.9	0	60.1
Node 10	Gene No	4199	1417	1040	1742	**Node 21**	Gene No	8391	3452	4057	882
	%	100	33.7	24.8	41.5		%	100	41.1	48.4	10.5

Tree scale: 0.1

Geobacillus stearothermophilus ATCC 12980 NZ JYNW01000038
Geobacillus thermodenitrificans DSM 465 NZ AYKT01000065
Geobacillus thermodenitrificans NG80-2
Geobacillus uzenensis BGSC 92A1 NZ NEWL00000000
Geobacillus subterraneus KCTC3922 NZ CP014342 1
Geobacillus subterraneus
Geobacillus proteiniphilus 1017 NZ MQMG00000000
Geobacillus sp. C56-T3 NC 014206
Geobacillus kaustophilus NBRC 102445 NZ BBJV01000028
Geobacillus sp. CAMR5420 NZ JHUS00000000
Geobacillus thermoleovorans SURF-48B NZ NAGD00000000
Geobacillus thermoleovorans
Geobacillus sp. LEMMJ02 NZ VKJO00000000
Geobacillus sp. Y412MC52 CP002442
Geobacillus thermocatenulatus BGSC 93A1 NEWK01000001
Geobacillus sp. GHH01 NC 020210
Geobacillus zalihae JS1-2 CP129217
Geobacillus jurassicus NBRC 107829 BCQG00000000
Geobacillus vulcani PSS1 NZ JPOI01000001
Geobacillus icigianus G1w1 NZ JPYA01000193
Geobacillus icigianus
Bacillus alveayuensis
Bacillus humi
Bacillus panaciterrae DSM 19096
Bacillus sp. FJAT-14578
Bacillus flexus NBRC 15715
Bacillus megaterium NBRC 15308 = ATCC 14581
Bacillus smithii
Bacillus methanolicus MGA3
Bacillus sp. UNC438CL73TsuS30
Bacillus dielmoensis
Bacillus sp. X1
Anoxybacillus tepidamans PS2
Anoxybacillus gonensis
Parageobacillus caldoxylosilyticus NBRC 107762
Parageobacillus toebii NBRC 107807
Parageobacillus thermoglucosidasius
Geobacillus lituanicus N-3 CP017692



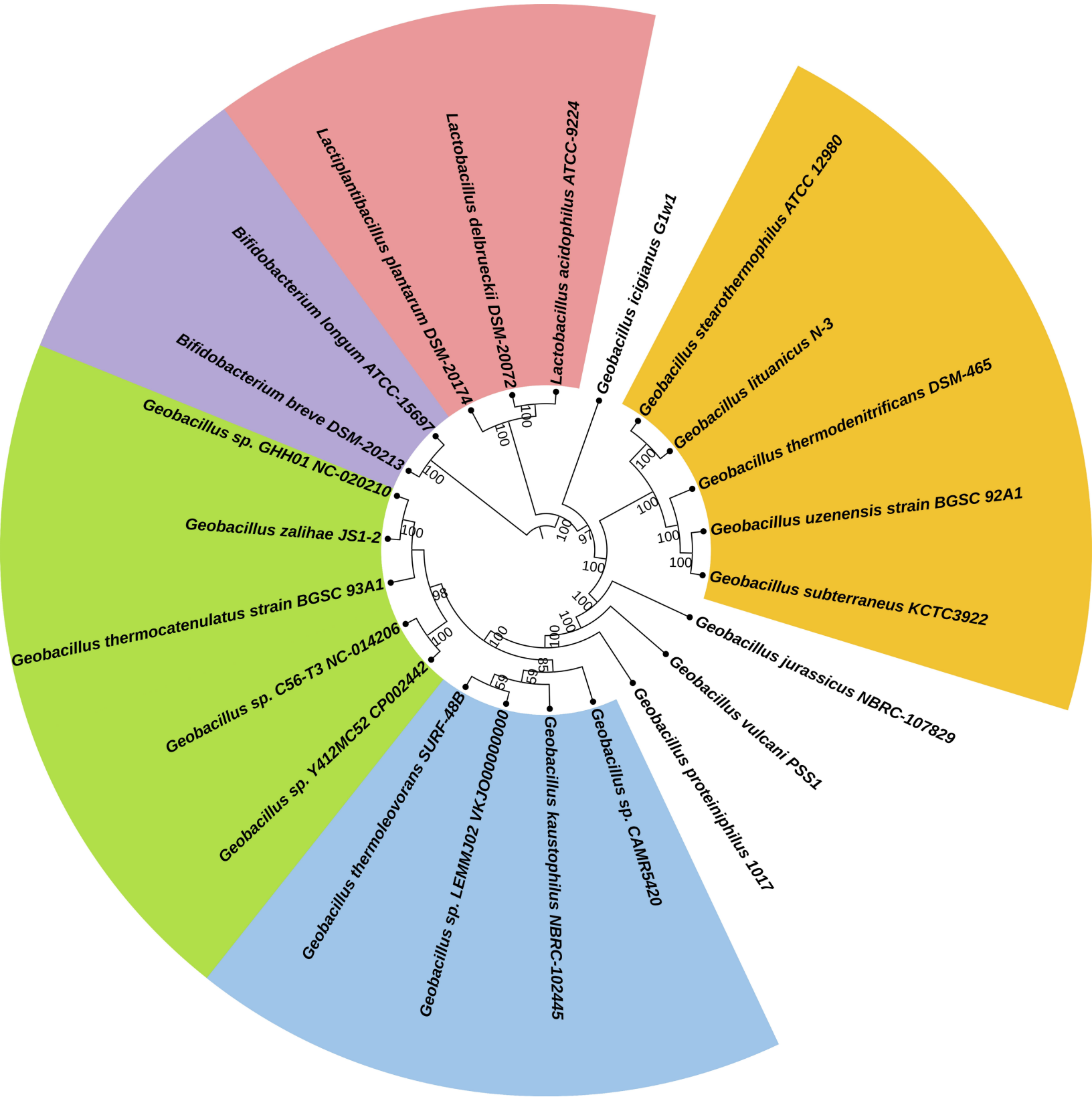


Table.1. Quality control, source and GC content of studied genomes.

Genomes	Source	Quast			CheckM	
		Total Length	N50	GC	Completeness	Contamination
***Geobacillus_icigianus*_G1w1__NZ_JPYA01000193**	Hot spring sediments	3457810	39998	52.03	99.45	0.34
***Geobacillus_kaustophilus*_NBRC_102445__NZ_BBJV01000028**	Pasteurized milk	3454242	50413	51.86	98.31	0
***Geobacillus_jurassicus*_NBRC_107829__BCQG00000000**	Unknown	3670957	3670957	52.22	99.45	0
***Geobacillus_lituanicus*_N-3__CP017692**	High-temperature oilfield	3499511	3447958	52.14	98.91	0.55
***Geobacillus_proteiniphilus*_1017__NZ_MQMG00000000**	High-temperature heavy oil reservoir	3574950	52882	51.76	99.45	0.68
***Geobacillus_stearothermophilus*_ATCC_12980__NZ_JYNW01000038**	Spoiled canned food	2860911	2831741	52.77	98.61	0
***Geobacillus_subterraneus*_KCTC3922__NZ_CP014342_1**	Oil field	3474426	3474426	52.2	99.45	0
***Geobacillus_thermocatenulatus*_BGSC_93A1__NEWK01000001**	Oil field	3563800	1841649	51.77	99.45	0.03
***Geobacillus_thermodenitrificans*_DSM_465__NZ_AYKT01000065**	Sugar beet juice	3400546	264274	49.05	99.45	0
***Geobacillus_thermoleovorans*_SURF-48B__NZ_NAGD00000000**	Yates Shaft, SURF	3808577	3808577	52.19	99.45	1.5
***Geobacillus_uzenensis*_BGSC_92A1__NZ_NEWL00000000**	Oil field	3357711	429603	52.23	99.45	0
***Geobacillus*_sp._GHH01__NC_020210**	Botanischer Garten	3583134	3583134	52.28	99.45	0
***Geobacillus*_vulcani_PSS1__NZ_JPOI01000001**	Unknown	3389115	3389115	52.4	99.45	1.39
***Geobacillus*_sp._CAMR5420__NZ_JHUS00000000**	Unknown	3499823	95605	51.89	99.45	0.14
***Geobacillus*_sp._Y412MC52__CP002442**	hot spring	3673940	3628883	52.34	99.45	0
***Geobacillus*_sp._C56-T3__NC_014206**	hot spring	3650813	3650813	52.49	99.45	0
***Geobacillus*_sp._LEMMJ02__NZ_VKJO00000000**	Fumarole sediment	3436274	150524	52.48	99.45	0.69

***Geobacillus_zalihae*_JS1-2__CP129217**	Compost	3567837	3532984	52.26	99.45	1.41
***Lactobacillus_acidophilus*_ATCC_9224__NZ_CP 130437**	Gastrointestinal tract	2035536	2035536	34.76	98.3	0
***Lactobacillus_delbrueckii*_subsp._*lactis*_DSM_200 72__CP022988**	Emmental cheese	2165984	2165984	49.05	100	0.75
***Lactiplantibacillus_plantarum*_DSM_20174__CP0 39121**	Pickled cabbage	3250154	3242936	44.5	99.07	2.78
***Bifidobacterium_breve*_DSM_20213__BBAO0000 0000**	Gastrointestinal tract	2232812	35183	58.69	100	0
***Bifidobacterium_longum*_subsp._*infantis*_ATCC_ 15697__CP001095**	Intestine of infant	2832748	2832748	59.86	100	0

Table.2. Prediction of genes related to probiotics

Function	*Genes*	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*O*	*P*	*Q*	*R*	*S*	*T*	*U*	*V*	*W*
Modulation of the immune system/Acid stress	***clpB***	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	P
	cysS	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	cysE	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	P	N	P
	cysK	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	N	P	P
	nrdD	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	P	P	P
	nrdR	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	nrdH	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	P	N	P
	nrd E	P	P	P	P	P	P	P	P	P	P	N	N	P	P	N	P	P	P	P	P	P	N	P
	nrdB	P	P	P	P	P	P	P	P	P	P	N	N	P	P	P	P	P	P	N	N	N	N	P
	nrdF	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	P	P	P
	pyrC	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	pyrE	P	P	P	P	P	N	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	pyrH	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	pyrR	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	P	P	P
	pyrP	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	P	P	P
	pyrB	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	pyrG	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	pyrK	P	P	P	P	P	P	N	P	P	P	P	P	P	P	P	P	P	P	P	P	N	P	P
	pyrF	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	pyrDB	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	N
	trxA	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	P	P	P
	trxB	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	yjbK	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	N	N	P
	yjbM	N	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	P	P	P
	dltC	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	P
Adhesion	***srtA***	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	srtC	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N

	srtD	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	N	N	N
	fbpA	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	lpsA	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	dltD	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	P
	dltA	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	P
	ClpB	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	P
	ClpX	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	ClpS	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	ClpC	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	strA1	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P
	strA2	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	strA3	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	tuf	P	P	P	P	P	N	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	lapA	N	P	N	N	P	N	N	N	N	N	P	N	N	P	N	P	P	P	N	N	N	N	N
	PilA	N	N	N	P	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	PilB	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	PilT	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	N	N	N
	PilC	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	PilM	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	PilN	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	PilO	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	PilZ	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	N	N	N	N	N	N
	comC	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	P	N	N
Bacterial colonization	***TadA***	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	TadB	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	TadC	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	TadE	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	TadF	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	TadZ	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
Aci d str ess	***atpC***	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	p	P

	atpD	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	atpG	P	P	P	P	P	P	P	P	P	P	N	P	P	P	P	P	P	P	P	P	P	P	P
	atpH	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	P	P	P
	atpF	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	atpB	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	atpE	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	P
	recA	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	sodA	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	N	N	P
	relA	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	aspS	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	cspA	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	N	P	N
Acid stress/bile resistance	*gpmA*	N	N	N	P	P	P	P	N	N	N	P	P	P	N	P	N	N	N	P	P	P	P	P
	gpml	N	P	N	N	P	P	P	P	P	P	P	P	P	P	P	N	P	P	N	N	N	N	N
	glmU	P	P	P	P	P	P	N	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	glmS	P	P	P	P	P	P	N	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	glmM1	P	P	P	P	P	N	N	N	N	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	glmM2	P	P	P	P	P	N	N	N	N	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	bshA	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	N	N	N
	bshB	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	N	N	N
	bshC	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	N	N	N
	luxS	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
Bile resistance	*nagB*	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	pyrG	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	argS	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	rpsC	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	rpsE	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	P	P	P
	rplD	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	rplE	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	rplF	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
Osmotic Stress	*opuA*	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N

	opuC	N	N	N	N	N	P	P	P	N	P	P	P	P	N	N	N	N	N	N	N	N	N	N
	opuD	P	N	P	N	P	P	P	P	P	N	N	N	N	N	N	P	P	N	N	N	N	N	N
	opuB	N	N	N	N	N	N	N	N	P	N	N	N	P	N	N	N	N	N	N	N	N	N	N
	opuCA	P	P	P	P	P	P	P	P	N	P	P	P	P	P	P	P	P	P	N	N	P	N	N
	oupCB	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	P	N	N
	opuCC	P	P	P	P	P	P	N	P	P	P	P	P	P	P	P	P	P	P	N	N	P	N	N
Molecular Chaperones	*groES*	N	N	N	P	N	N	N	N	N	N	N	N	N	N	P	N	P	N	N	N	N	N	N
	groEL	N	N	N	P	N	N	N	N	N	N	N	N	N	N	P	P	P	N	N	N	N	N	N
	dnaK	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	dnaJ	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	dnaG	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
Fatty acid synthesis	*fabD*	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	P	N	P
	fabH	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	P	N	P
	fabF	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	P	N	P
	fabL	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	N	N	N
	fabZ	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	P	N	P
	accC	N	P	N	N	P	N	N	N	P	P	P	P	N	P	P	P	P	P	N	N	P	N	P
Lactate synthesis	*mdh*	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	N	N	N
Others	*dapA*	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	dps	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	P
Transcriptional regulator	*sigB*	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	P	N	N	N
	sigX	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	P	N	N	N
	ctsR	P	P	P	P	P	P	N	P	P	P	P	P	P	P	P	P	P	P	N	N	P	N	N
	hrcA	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
Metabolic rearrangement	*alsD*	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	N	N	N	N	N	N	N
Vitamin Biosynthesis	*copA*	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	P	P	P
	copZ	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	P	P	P	N	N	N	P	N
	copZA	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	csoR	P	P	P	P	P	P	N	P	P	P	P	P	P	P	P	P	P	P	N	N	N	N	N
	BtuF	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	N	N	N

	BtuC	N	N	N	P	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	BtuD	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
Oxidative Stress	***trxA***	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	P	P	P
	trxB	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
	feoB	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	P	N	P
	oxyR	N	N	N	N	N	N	N	P	P	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	cydA	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	cydC	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	yocS	N	N	N	P	N	N	N	N	N	N	N	N	N	N	P	N	N	N	N	N	N	N	N
	gpx	N	N	P	P	N	N	N	P	N	N	N	P	N	N	P	N	N	N	N	N	N	N	N
	ndh	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	N	N	N
	ndhF	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	N	N	N	N	N	N	N	N
	ndhS	P	P	N	P	P	P	P	N	P	P	P	P	P	P	P	P	P	P	N	N	N	N	N
	ndhB	P	N	P	N	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	N	N	N	N
	ndhC	P	P	P	P	P	P	N	P	P	P	P	P	P	P	P	P	P	P	N	N	N	N	N
	ndhH	P	P	P	P	P	N	N	N	N	N	N	N	N	P	P	P	P	P	N	N	N	N	N
	ndhD	N	N	N	P	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	msrA	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	P	P	P
	msrB	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	P
	msrC	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	P	P	P
CRISPR-Cas	***Cmr1***	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	N	N	N	N	N	N	N
	Cmr2	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	N	N	N	N	N	N	N
	Cmr3	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	N	N	N	N	N	N	N
	Cmr4	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	N	N	N	N	N	N	N
	Cmr5	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	N	N	N	N	N	N	N
	Cmr6	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	N	N	N	N	N	N	N
	TM1812	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	Csd1	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	Csd2	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	Cas5h	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	N	N	N	N	N	N

	Cas3	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	N	N	N	N	N	N
	Cas4	N	N	N	N	N	N	N	N	N	N	N	N	P	N	N	N	P	N	N	N	N	N	N
	Cas1	P	N	P	N	N	P	P	P	N	N	P	P	P	N	N	P	P	P	N	N	N	N	N
	Cas2	P	N	P	N	N	P	P	P	P	N	P	P	P	N	N	P	P	P	N	N	N	N	N
	Cas6	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	P	N	N	N	N	N	N
	cas9	P	P	N	N	N	N	P	N	N	N	N	P	N	N	N	N	N	P	N	N	N	N	N
	cas10	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	N	N	N	N	N	N	N	N
	csx1	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	N	N	N	N	N	N	N
	csm6	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	N	N	N	N	N	N	N
	cst1	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	N	N	N	N	N	N
	DevR	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	N	N	N	N	N	N

A= *G. icigianus*, B= *G. thermoleovorans*, C= *G. uzenensis*, D= *G. vulcani*, E= *G. zalihae*, F= *G. lituanicus*, G= *G. stearothermophilus*, H= *G. subterraneus*, I= *G. thermodenitrificans*, J= *G. kaustophilus*, K= *G. proteiniphilus*, L= *G. jurassicus*, M= *G. thermocatenulatus*, N= *G. sp. CAMR5420*, O= *G. sp. GHH01*, P= *G. sp. Y412MC52*, Q= *G. sp. C56-T3*, R= *G. sp. LEMMJ02*, S= *B. breve*, T= *B. longum*, U= *L. plantarum*, V= *L. acidophilus*, W= *L. delbrueckii*.

P=Positive/Present, N=Negative/Not Present

Table.3. ARGs, Plasmids and Virulence factors present in studied genomes.

Genomes	CARD PATRIC	CARD PROKSEE	CARD	PLASMIDS	VF (Patric)	VF (Abricate)
G. icigianus	0	*vanT, vanY*	*vanT, vanY*	0	0	0
G. thermoleovorans	0	*vanT, vanY*	*vanT, vanY*	0	0	0
G. uzenensis	0	*vanT, vanY*	*vanT, vanY*	0	0	0
G. vulcani	0	*vanT, vanY*	*vanT, vanY*	0	0	0
G. sp. CAMR5420	0	*vanT, vanY*	*vanT, vanY*	0	0	0
G. sp. GHH01	0	*vanT, vanY*	*vanT, vanY*	0	0	0
G. sp. Y412MC52	0	*vanT, vanY*	*vanT, vanY*	0	0	0
G. sp. C56-T3	0	*vanT, vanY*	*vanT, vanY*	0	0	0
G. sp. LEMMJ02	0	*vanT, vanY*	*vanT, vanY*	0	0	0
G. zalihae	0	*vanT, vanY*	*vanT, vanY*	0	0	0
G. jurassicus	0	*vanT, vanY*	*vanT, vanY*	0	0	0
G. kaustophilus	0	*vanT, vanY*	*vanT, vanY*	0	0	0
G. lituanicus	Linzolid	*vanT, vanY*	*vanT, vanY*	0	0	0
G. proteiniphilus	0	*vanT, vanY*	*vanT, vanY*	0	0	0
G. stearothermophilus	0	*vanT, vanY*	*vanT, vanY*	0	0	0
G. subterraneus	0	*vanT, vanY*	*vanT, vanY*	0	0	0
G. thermocatenulatus	0	*vanT, vanY*	*vanT, vanY*	0	0	0
G. thermodenitrificans	0	*vanT, vanY*	*vanT, vanY*	0	0	0
B. breve	*rrsA, ileS, rpoB*	*rpoB*	*rpoB*	0	0	0
B. longum		*rpoB, rpsL*	*rpoB, rpsL*	0	0	0
L. plantarum		*vanH, vanY*	*vanY*	0	0	0
L. acidophilus		*vanT*	*vanT*	0	0	0
L. delbrueckii		*ImrD, vanY, qacJ*	*ImrD, vanY, qacJ*	0	0	0

Table.4. Presence of genes related to Toxins.

Gene_ Description	HMM_ id	G 1	G 2	G 3	G 4	G 5	G 6	G 7	G 8	G 9	G1 0	G1 1	G1 2	G1 3	G1 4	G1 5	G1 6	G1 7	G1 8	B 1	B 2	L 1	L 2	L 3
PhoH-like Protein	T.21348	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	P	P	P	P	P
Bacterial extracellular solute-binding protein, family 7	T.10571	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	N	N	P	N	N
PemK-like, MazF-like toxin of type II toxin-antitoxin system	T.56898	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	P	P	P	N	N
Fic/DOC family N-terminal	T.42972	N	N	N	P	N	N	P	P	P	P	N	P	P	P	P	P	P	N	N	N	N	N	N
Nucleotidyl transferase AbiEii toxin, Type IV TA system	T.13738	N	N	N	N	P	P	P	P	N	N	P	N	N	P	N	P	P	N	N	N	N	N	N
Fic/DOC family	T.45183	P	N	P	P	P	P	P	P	N	P	N	P	P	P	P	P	P	N	N	P	N	P	N

HicA toxin of bacterial toxin-antitoxin,	T.27346	N	N	N	N	P	P	N	N	N	P	P	N	N	N	N	N	N	N	N	N	N	N	N
TRAM domain	T.44171	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P
HipA-like C-terminal domain	T.39963	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	N	N	N	N
ParB-like nuclease domain	A.66515	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	P	P	P
RelB antitoxin	A.56826	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	N	P	N
Replication initiation factor	A.60023	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	N	N	N	N
HipA-like C-terminal domain	T.59740	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	N	N	N
MraZ protein, putative antitoxin-like	A.62916	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	P	P	N
Phage derived protein Gp49-like (DUF891)	T.41827	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	N	N	N	N
Bacterial	T.4367	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	N	N	P	N

toxin of type II toxin-antitoxin system, YafQ	0																							
PIN domain	T.15594	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	P	N	N	N
YoeB-like toxin of bacterial type II toxin-antitoxin system	T.13352	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	N	N	N
RelE-like toxin of type II toxin-antitoxin system HigB	T.9856	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	P	N	N

G1= *G. icigianus*, G2= *G. jurassicus,* G3= *G. kaustophilus,* G4= *G. lituanicus,* G5= *G. proteiniphilus,* G6= *G. stearothermophilus*, G7= *G. subterraneus,* G8= *G. thermocatenulatus*, G9= *G. thermodenitrificans,* G10= *G. uzenensis*, G11= *G. vulcani*, G12= *G. sp. CAMR5420,* G13= *G. sp. LEMMJ02,* G14= *G. sp. Y412MC52*, G15= *G. sp. C56-T3,* G16= *G. sp. GHH01*, G17= *G. zalihae*, G18= *G. thermoleovorans,* B2= *B. breve*, B1= *B. longum*, L1= *L. plantarum*, L2= *L. acidophilus*, L3= *L. delbrueckii*.

P=Positive/Present, N=Negative/Not Present

Table.5. Presence of bacteriocins in studied genomes.

Genome	Name	Funtion	Motif
G. icigianus	Circularin_A	Bacteriocin_IId; 147.1; Circularin_A	PF09221
	318.1; ComX1	318.1; ComX1	
	Sactipeptides	Uncharacterized MFS-type transporter YfkF OS	
G. jurassicus	0		
G. kaustophilus	37.1; Geobacillin_I_like	MA-NIS+EPI; MA-NIS; 37.1; Geobacillin_I_like	RBS=G
G. lituanicus	37.1; Geobacillin_I_like	leaderLanBC; LE-LanBC; MA-EPI; MA-NIS+EPI; MA-NIS; Gallidermin; 37.1; Geobacillin_I_like	PF02052
	107.1; Salivaricin_D	107.1; Salivaricin_D	
G. proteiniphilus	0		
***G. sp.* C56-T3**	321.1; ComX4	321.1; ComX4	
	147.1; Circularin_A	Bacteriocin_IId; 147.1; Circularin_A	PF09221
***G. sp.* CAMR5420**	321.1; ComX4	321.1; ComX5	
***G. sp.* GHH01**	0		
***G. sp.* LEMMJ02**	0		
***G. sp.* Y412MC52**	147.1; Circularin_A	Bacteriocin_IId; 147.1; Circularin_A	PF09221
G. stearothermophilus	147.1; Circularin_A	Bacteriocin_IId; 147.1; Circularin_A	PF09222
G. subterraneus	0		
G. thermocatenulatus	37.1; Geobacillin_I_like	MA-NIS+EPI; MA-NIS; 37.1; Geobacillin_I_like	RBS=TG
	147.1; Circularin_A	Bacteriocin_IId; 147.1; Circularin_A	PF09221
G. thermodenitrificans	36.1; geobacillin_I	leaderLanBC; MA-EPI; MA-NIS+EPI; MA-NIS; Gallidermin; 36.1; geobacillin_I	PF02052
G. thermoleovorans	37.1; Geobacillin_I_like	LE-LanBC; 37.1; Geobacillin_I_like	

G. uzenensis	0		
G. vulcani	490.1; Pumilarin	Bacteriocin_IId; 490.1; Pumilarin	PF09221
G. zalihae	321.1; ComX4	321.1; ComX4	
B. longum	37.1; Geobacillin_I_like	leaderLanBC; MA-EPI; MA-NIS+EPI; MA-NIS; Gallidermin; 37.1; Geobacillin_I_like	PF02052
	7.1; BLD_1648	Mersacidin; 7.1; BLD_1648	PF16934
	186.2; Propionicin_SM1	186.2; Propionicin_SM1	
B. breve	0		
L. plantarum	170.2; Plantaricin_E	170.2; Plantaricin_E	
	171.2; Plantaricin_F	ggmotif; Lactococcin; Bacteriocin_IIc; 171.2; Plantaricin_F	PF04369; PF10439
	167.2; Plantaricin_A	Antimicrobial17; Bacteriocin_IIc; 167.2; Plantaricin_A	PF08129; PF10439
	174.2; Plantaricin_N	Bacteriocin_IIc; 174.2; Plantaricin_N	PF10439
	172.2; Plantaricin_J	172.2; Plantaricin_J	
	173.2; Plantaricin_K	ComC; Bacteriocin_IIc; 173.2; Plantaricin_K	PF03047; PF10439
L. acidophilus	97.2; Enterocin_X_chain_beta	97.2; Enterocin_X_chain_beta	
	6.2; Acidocin_J1132_beta_peptide_N-terminal	Bacteriocin_IIc; 6.2; Acidocin_J1132_beta_peptide_N-terminal	PF10439
	64.3; Enterolysin_A	64.3; Enterolysin_A	
	6.3; Bacteriocin_helveticin_J	6.3; Bacteriocin_helveticin_J	
L. delbrueckii	63.3; Enterolysin_A	63.3; Enterolysin_A	

Table.6. Presence of Prophages in studied genomes.

Genomes	Scaffold	Begin	End	Transposable	Taxonomy
G. icigianus	GCA_000750005.1-30	2666	13563	FALSE	Myoviridae
G. thermoleovorans	GCA_001610955.1	2273102	2279397	FALSE	Myoviridae / Siphoviridae
	GCA_001610955.1	3090580	3093683	TRUE	Unknown
G. uzenensis	0				
G. vulcani	GCA_000733845.1	272535	293893	FALSE	Siphoviridae
	GCA_000733845.1	296023	305288	FALSE	Siphoviridae
	GCA_000733845.1	1473929	1477744	FALSE	Unknown
G. sp. CAMR5420	0				
G. sp. GHH01	CP004008.1	798882	806496	TRUE	Myoviridae
G. sp. Y412MC52	CP002442.1	2966755	2991656	FALSE	Siphoviridae
	CP002442.1	2993517	3001469	FALSE	Unknown
G. sp. C56-T3	CP002050.1	256841	291284	FALSE	Siphoviridae
G. sp. LEMMJ02	0				
G. zalihae	0				
G. jurassicus	Ga0128515_1001	35934	47638	FALSE	Myoviridae / Siphoviridae
	Ga0128515_1027	19546	45535	FALSE	Myoviridae
G. kaustophilus	GCA_000739955.1-78	2021	35631	FALSE	Myoviridae
G. lituanicus	GCA_002243605.1	2236888	2243431	TRUE	Unknown
G. proteiniphilus	Ga0347354_038	2748	33525	FALSE	Myoviridae
G. stearothermophilus	GCA_001277805.1-68	3732	10749	FALSE	Siphoviridae
	GCA_001277805.1-68	29604	41176	FALSE	Myoviridae / Siphoviridae
G. subterraneus	0				
G. thermocatenulatus	0				
G. thermodenitrificans	0				
B. longum	CP059048.1	534111	543342	FALSE	Siphoviridae
	CP059048.1	1707977	1741362	FALSE	Siphoviridae

B. breve	AP012324.1	1485228	1507366	FALSE	Siphoviridae
L. plantarum	CP025991.1	1732	21277	FALSE	Siphoviridae
	CP025991.1	692782	722751	FALSE	Siphoviridae
	CP025991.1	1192860	1224623	FALSE	Siphoviridae
	CP025991.1	1318309	1338974	FALSE	Siphoviridae
	CP025991.1	1345766	1349279	FALSE	Siphoviridae
	CP025991.1	1714169	1737434	FALSE	Siphoviridae
	CP025991.1	1742814	1749863	FALSE	Siphoviridae
L. acidophilus	0				
L. delbrueckii	CP059048.1	534111	543342	FALSE	Siphoviridae
	CP059048.1	1707977	1741362	FALSE	Siphoviridae