Genomic Features and Comparative Genomic Analysis of novel *Bacillus glycinifermentans* strain JRCGR-1

Asad Karim (asad.karim.alvi@gmail.com)
International Center for Chemical and Biological Sciences
poirot olivier
Aix Marseille University, UMR 7256
Ambrina Khatoon
Ziauddin University
Matthieu Legendre
Aix Marseille University, UMR 7256

Research Article

Keywords: B. glycinifermentans sp, pangenome, comparative genomic, Bacillus sp

Posted Date: September 16th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-902697/v1

Abstract

To the best of our knowledge, only six *B. glycinifermentans* sp. genome sequences are available in the public database. Here, we performed genome sequencing and comparative genomics analysis of *B. glycinifermentans* strain JRCGR-1. Cluster analysis of strain JRCGR-1 genes showed that 92.6% of genes were present in the orthogroups and 7.4% genes were not assigned to any group. The pangenome size was calculated at 8329 genes and presented an open genome characteristic. Phylogeny based on the pan and core genome demonstrated that all the *B. glycinifermentans* strains belong to the same clade. The strain JRCGR-1, ANI, TETRA and DDH values were in the range of 96.1-99.04%, 0.996-997, 73.5–84.7%, respectively. The strain JRCGR genome exhibits a high level of *synteny* with multiple locations in *B*. sonorensis sp. and *B. licheniformis* sp. The finding of the current study provides knowledge that facilitates a better understanding of this at the genomic level.

Introduction

Member of the genus *Bacillus* are ubiquitous in the environment, and they have a huge impact on human activities. *Bacillus* sp. are rod-shaped, spore-forming, gram-positive and aerobic bacteria [1]. More than 200 *Bacillus* sp. have been identified and their biochemistry and physiology have been studied. For the taxonomy of microorganisms, the 16S rRNA gene is frequently. However, for highly similar microorganisms 16S rRNA genes cannot be used. For instance, the *B. pumilus* group are distinguished through the gyrB (β-subunit of DNA gyrase) gene, chemotaxonomic properties phenotypic characteristics and DNA-DNA connectedness [2, 3]. Therefore, bacteria are now reclassified due to the advancement in phylogenetic analysis, DNA-DNA hybridization and other molecular techniques [1]. Specifically, comparative genomic analysis proved valuable information for studying the taxonomy of microorganisms because comparative genomics true essence is linked to how we explore the species evolutionary relationships. Evolutionarily linked genes (homologs) are either genes that originated from speciation events (orthologs) or gene duplication events (paralogs) [4]. This distinction is important for broader range analysis, such as phylogenetic tree building, comparative genomics, genome annotation and prediction of gene function [5–7].

Since ancient times, *Bacillus* sp. have been used for fermentation and they are the main workhorses in applied microbiology. The genus has many species which significant because of their specific properties such as biofilms formation, antibiotics activities, probiotics and production of enzymes. In 2015, a novel *Bacillus* sp. named *B. glycinifermentans* was first isolated from a Korean fermented soybean paste food (cheonggukjang) [8]. To the best of our knowledge, we have reported the first draft genome sequence of *B. glycinifermentans* sp. from Pakistan. We previously published a draft genome sequence of *B. glycinifermentans* strain JRCGR-1 (Assembly number, VHPY00000000) [9]. Including strain JRCGR-1, only six *B. glycinifermentans* genome sequences are available in the NCBI database; three complete genome sequences and three draft genomes genome sequences

In the current study, we systematically used several methods to comprehensively analyze and compare the genomic features of strain JRCGR-1 and its closely related species. First, we analyzed the evolutionary relationship of this strain with its closely related *Bacillus* sp. based on 16sRNA, ANI, AAI, TETRA and codon usage. Second, we analyse the pangenome, orthology and synteny analyses. Third, we performed functional annotation of core, accessory and unique genes. Finally, to explore its commercial importance, we performed annotation of strain JRCGR-1 proteome and several putative genes were predicted that can produce industrially important enzymes.

Methods

Strain Isolation

The strain was isolated from a soil sample in Karachi, Pakistan. For isolation, the soil (1 g) sample was serially diluted with normal saline water up to 10^{-10} dilution, and from the last four dilutions (10^{-7}, 10^{-8} 10^{-9} and10^{-10}), 40 µl samples were transferred to nutrient agar plates. After incubation for 24hrs. at 37 ºC, single colonies were picked and their characteristic was noted, followed by Gram staining.

Genome sequencing, assembly, and annotation

We have previously reported the genome sequence, assembly and annotation of strain JRCGR-1[9]. In brief, a colony of bacteria was inoculated in Luria broth for 24 h at 37 ºC. The optical density was adjusted to a McFarland standard (3-4nm). After incubation, bacterial cells were harvested by centrifugation and DNA extraction was performed by using a commercially available kit (QIAamp1 DNA Mini Kit; QIAGEN, Hilden, Germany). The amount of DNA was calculated on a Qubit1 2.0 fluorometer (Invitrogen) using a QubitTM dsDNA BR Assay kit (Invitrogen; Thermo Fisher Scientific, Eugen, OR, USA) and the paired-end library was constructed using a Nextera XT DNA Library Kit (Illumina Inc., San Diego, CA, USA). The genome sequencing was performed using an Illumina NextSeq 500 platform (Illumina, San Diego, CA USA). The quality of the reads was elevated using FastQC v0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The quality of the reads was improved by trimming using Sickle and the DNA short reads were assembled using SPAdes (v3.12.0). Finally, 83 scaffolds and 84 contigs were generated for strain JRCGR-1. The quality of contigs or scaffolds were tested by QUAST 5.02.[10] and the gene annotation was performed using Prokka v.1.13 [11]. tRNA genes were identified by tRNAscan-SE [12]. RNAmmerr and Barrnap v0.4.2 were used to find out RNA genes [13]. We also assemble and annotated the Plasmid of strain JRCGR-1 using plasmidSPAdes [14] and Prokka (version v.1.13) software [11], respectively.

Bacillus species in this study

At the time of writing of this paper, three whole-genome sequences were available in the GenBank and two draft genome sequence was available in the NCBI database. To study the phylogenetic origin of strain JRCGR-1, we used three complete genome sequences of *B. glycinifermentans* sp. three closely

related *Bacillus* sp. (*B. licheniformis*, *B. paralicheniformis*, *B. sonorensis*) and two outgroup *Bacillus* sp. (*B. velezensis* and *B. subtilis*) were included in the study (Table S1).

16s rRNA sequencing and phylogenetic analysis

B. glycinifermentans JRCGR-1 partial 16S ribosomal RNA gene sequence was retrieved from the draft genome and blast against the NCBI nucleotide database. These sequences were aligned using muscle function in the MEGA X software [4] and the evolutionary history was inferred using the UPGMA method [15]. The optimal tree with the sum of branch length = 1.41530778 is shown with optimal tree. The evolutionary distances were calculated by the Maximum Composite Likelihood method [16] and are in the units of the number of base substitutions per site. This analysis involved 21 nucleotide sequences. For each sequence pair, all ambiguous positions were removed. In the final data set, there were a total of 1742 positions. FigTree v1.4.4 was used for the graphical viewer of the phylogenetic tree (http://tree.bio.ed.ac.uk/software/figtree/).

Multilocus Sequence Analysis (MLSA)

The gene sequences of gyrA, gyrB, rpoB and aptD retreated from the complete genome sequence of *B. glycinifermentans* JRCGR-1. These sequences were separately BLAST against the GenBank database (until November 2019), to find out the homologues with the highest sequence identity. The tree was generated from gyrA, gyrB, rpoB and aptD gene sequences. The UPGMA method was used to inferred evolutionary history [15]. The sum of branch length = 0.06553352 is shown with optimal tree. Poisson correction method was used to calculate the evolutionary distances [17] and its unit was in amino acid substitutions per site. Five amino sequences were used for the analysis and each sequence pair (pairwise deletion option), all ambiguous positions were removed. A total of 473 positions were in the final dataset. MEGA X [4] software was used for evolutionary analyses and Figtree 1.4.3 was used to visualized trees. (http://tree.bio.ed.ac.uk/software/figtree/)

Species identification using whole-genome phylogenetic analysis

To find out the phylogenetic origin with respect to other *Bacillus* sp., the whole genome sequences of seventeen publicly published *Bacillus* sp. presenting sequence identity similarity with *B. glycinifermentans* JRCGR-1 were used for comparative genomics. These *Bacillus* sp. included, *B. glycinifermentans*, *B. licheniformis*, *B. paralicheniformis*, *B. sonorensis* and two outgroup *Bacillus* sp. (*B. velezensis* and *B. subtilis*) were included in the study (Table S1).

CompareM v0.0.24 software toolkit was to find out pairwise genome statistics (e.g., amino acid identity (AAI)) and statistics for individual genomes (e.g., codon usage,) (https://github.com/dparks1134/CompareM). DIAMOND-0.814 software was used for similarity search (https://github.com/bbuchfink/diamond) and Prodiagal-2.6.3 (https://github.com/hyattpd/Prodigal) was used for gene calling over the genome. For taxonomic classification, the query (*B. glycinifermentans*

JRCGR-1) genome was compared to the target genomes (selected 19 *Bacillus* sp. given in Table S1) using CompareM v0.0.24 software and the putative homologs were identified.

Genome-to-Genome Distance Calculator (GGDC)(calculated by formula 3) was used to calculate the in *silico* DDH (DNA-DNA hybridization) values [18]. We used Jspecies software [19] to compute the tetranucleotide signatures (TETRA) and the average nucleotide identity (ANI) values. For ANI, we used NUCmer (ANIm) [20] algorithms. All the heat maps were plotted using the R package [21].

General genome analyses

The pan-genome and core genome were identified from orthologous gene clusters; OrthoFinder [22] output file was converted to PanGP [23] format for pan-genome analysis. We also constructed a phylogenetic tree using core and pan genes with Bacterial Pan Genome Analysis (BPGA) software [24]. Figtree 1.4.3 was used to visualize the Newick files. (http://tree.bio.ed.ac.uk/software/figtree/).

Orthology and Synteny analyses

Orthologous genes between *Bacillus* sp. was analyzed by OrthoFinder [22]. For the synteny analysis, we used *B. glycinifermentans* JRCGR-1 as query and three strains (*B. glycinifermentans* and *B. licheniformis* and *B. sonorensis*) were used as a subject. First, OrthoFinder [22] was used to find out the orthologous genes between query and subject. Second, iAdhore [25] was used to identify longer-term ancestral synteny and finally, circus [26] was used to draw synteny plots.

Online service (https://orthovenn2.bioinfotoolkits.net/home) was used to draw a Venn diagram for the distribution of shared gene families (orthologous clusters) among *B. glycinifermentans* six strain, including JRCGR-1, BGLY, KBN06PO3352, SRCM103574, KJ17 and GO13 (Table S2).

Submitting nucleotide sequences

The draft genome sequence of *B.* glycinifermentans JRCGR-1 was submitted to the NCBI database with accessibility number VHPY00000000 [9].

Results And Discussion

Species confirmation by 16S rRNA gene and MLSA

NCBI blast results show that *B. glycinifermentans* JRCGR-1 16S ribosomal RNA gene has 99.91% similarity with *B. glycinifermentans* BGLY (NZ_LT603683.1) and *B. glycinifermentans* SRCM103574 (NZ_CP035232.1). In the phylogenetic tree, both strains SRCM103574 and BGLY were present in the same branch (Fig. 1A). However, *B. glycinifermentans* JRCGR-1 was present in the same clade but as a single branch. Five strains of *B. glycinifermentans* (highlighted in blue) were present as two separate clades

which were close to each other. *B. licheniformis* sp. was the closest, and in contrast, *B.* velezensis and *B.* subtilis sp. were farthest away from JRCGR-1 based on 16S RNA gene sequence analysis.

Fig. 1

As for strain, BGLY, SRCM103574, KJ-17-KR092216.1, KBN06P03352 and JRCGR-1, sequence identity with any species contained in the database was above 97%, suggesting that strain JRCGR-1 could represent as *B. glycinifermentans* sp. [27].

To further confirm the results of the 16S rRNA gene sequence analysis, we analyzed four housekeeping genes of *B. glycinifermentans* sp.: gyrA (Fig. 1B), gyrB (Fig. 1C), rpoB (Fig. 1D), and aptD (Fig. 1E). Each one of these genes was BLAST against the GenBank protein database (NCBI) [28]. The finding reveals that all the housekeeping genes of strain JRCGR-1 presented higher similarities (≥97%) with other *B. glycinifermentans* strains in the GenBank database. These results confirm that JRCGR-1 is classified as a *B. glycinifermentans* sp.

Characteristic features of strain JRCGR-1

The strain JRCGR-1 genome contained 4700692 bp with 45.5%. average G+C content and the final assembly contains 84 contigs and 83 scaffolds [9]. The annotation results revealed 5174 genes, 32 tRNA, 4 rRNA, 1 tmRNA and 92 misc_RNA. PlasmidSPAdes was used to assembled plasmid into 37 contigs. The plasmid size was found to be 1113267 bp and it contains tRNA: 27, rRNA: 4, gene: 1366, misc_RNA: 21, CDS: 1314 [9].

Whole-genome sequence comparisons

ANI (ANIm), TETRA, AAI, codon usage and DDH values were calculated for species identification. The strain JRCGR-1 ANIm (Fig. 2A), TETRA (Fig. 3A) and DDH (Fig. 3B) values with respect to three strains of *B. glycinifermentans* sp. (BGLY, KBNO6P03352 and SRCM103574) were in the range of 96.1-99.04%, 0.996-997, 73.5-84.7%, respectively. The strains of *glycinifermentans* sp. used in the current study are within the region of boundaries defined for genomic species;0.99 for TETRA, 70% for DDH and 95–96 for ANI [19]. These values qualify strain JRCGR-1 as members of a single genomic species and are different from any other *Bacillus* sp.

Fig. 2

Heat map analysis of AAI (Fig 2B) also shows that strain JRCGR-1 is *glycinifermentans* sp.

Among twenty strains of *Bacillus* sp., *Bacillus* sp. H15 (NZ_CP018249) and *Bacillus* sp. 1S-1 (NZ_CP022874) were not previously classified in any species. Our study (based on ANIb, ANIm, TETRA, DDH and AAI) shows that they should be considered as a *B. licheniformis* sp. from a genomic point of view.

Fig 3

Orthology analysis

Cluster analysis of the selected *Bacillus* sp. showed that 97.83% of genes were in orthogroups and only 2.2% of genes were unassigned to any group. Total orthogroups were found to be 6403 and the mean orthogroups size value was 13.5. The overall statics is given in Table S3.

Phylogenetic tree of selected *Bacillus* sp. based on ortholog clustering shows that strain JRCGR-1 and BGLY are closely related strains (Fig. 4A). It is interesting to note that, the highest number of genes were present in strain JRCGR-1, which was followed by *B.* sonorensis strain SRCM10139 and three other *B.* glycinifermentans species (Fig. 4B). For each *Bacillus* sp., more than 90% of the genes were present in the orthogroups and the number of orthogroups containing species was found to be greater than 60% (Fig. 4C). Moreover, very few genes were unassigned to any orthogroups. Specifically, for strain JRCGR-1, out of 5045 genes, 4670 genes were present in the orthogroups (92.6%) and 375 genes were not assigned to any group (7.4%) (Fig. 4B and 4C). Fig. 4D shows several genes in species-specific orthogroups and the number of species-specific orthogroups.

Fig. 4

Synteny analysis

Synteny allows us to study the evolution between genomes, functional conservation [29-31], detect genome rearrangements [31], aid genome annotation [32] and helps to detect the quality of the genome assembly. In the current study, we illustrate a comparison featuring four primate genomes; strain JRCGR-1 was used as reference genome and *B*. sonorensis strain SRCM101395 and *B.* licheniformis strain ATCC 14580 as a query genome. Fig. 6 clearly shows that there is significant chromosomal synteny between the reference genome and the query genomes. For instance, total out of 83 contigs of stain JRCGR-1, 57 contigs shared synteny blocks with stain *B. glycinifermentans* BGLY genomes (Fig. 5A). Similarly, the reference genome also exhibits a high level of *synteny* with multiple locations in *B*. sonorensis strain SRCM101395 and *B. licheniformis* strain ATCC 14580 genomes (Fig. 5B and Fig. 5C). But, in the case of *B. licheniformis* strain ATCC 14580, the arrangement of synteny was different with reference to *B. glycinifermentans* BGLY and *B*. sonorensis strain SRCM101395 genomes.

These results indicate that the core genomes of the *B. glycinifermentans* strains are conserved and *B*. sonorensis is the closest species to *B. glycinifermentans* sp.

Fig. 5

Pan and core gene

Pangenome analysis has been used for the evaluation of the genome diversity, species evolution, pathogenesis and other characters of microorganisms [33]. Therefore, to better understand the bacterial evolution and the phylogenetic relationship, we analyzed the *Bacillus* sp. pangenome.

Fig. 6A shows that the number of new genes does not converge to zero upon sequencing of additional genomes. The future rate of the discovery of novel genes in a species can be predicted by analyzing the increase/decrease in the pangenome size with the addition of a new genome sequence [34, 35]. Fig. 6B shows the evolution of the pan and core genomes. The pangenome size is computed at 8329 genes (n = 20) and shows an open genome characteristic: (i) with the addition of genomes, the trajectory of the pangenome increases unboundedly and (ii) Bpan was calculated as 0.16. In the pangenome, core and accessory genes account for 30.77% and 69.10%, respectively.

Phylogeny based on the pan (Fig. 6C) and core (Fig. 6C) genome demonstrated that all the *B. glycinifermentans* strains belong to the same clade and *B.* sonorensis strain SRCM101395 is closer to *B. glycinifermentans* species. This finding reveals a lack of diverse genetic evolution of the pan and core genome in the four-strain of *B. glycinifermentans* sp.

Fig. 6

Figure 7, showing the heat map and the phylogenetic tree created by hierarchical clustering using Manhattan distance matrix based on the presence (red) and absence (blue) genes. Clade I include species of *B*. subtilis and *B*. velezensis, which exhibits high diversity at genome level with reference to Clade II and Clade III. The genomes in clade III show less diversity compared to clade I and clade II.

Fig. 7

Comparison among *B*. glycinifermentans

We also computed the orthologues genes for six strains of *B*. *glycinifermentans* sp. including JRCGR-1, KBN06PO3352, BGLY, SRCM103574, KJ17 and GO13 (Fig.8). Fig. 8A shows the graphical representation of orthologues genes shared between strain JRCG-1 and other *B*. *glycinifermentans* strains; from the outer circle inwards; 83 scaffolds of strain JRCG-1 (1); genes (5074) of strain JRCG-1 (2); orthologues genes (4670) assigned to strain JRCG-1 (3); orthologues genes (4196) shared between JRCG-1 and BGLY strain (4); orthologues genes (4018) shared between JRCG-1 and KBN06PO3352 strain (5) orthologues genes (4274) shared between JRCG-1 and SRCM103574 strain (6). unassigned genes (375) of strain JRCG-1 to any orthogroups (7).

A total of 3148 genes were shared by these six strains and only 30 genes were strain-specific (Fig. 8B). Thirteen strain-specific genes were found for strain JRCG1, five for strain SRCM103574, three for strain BGLY, one for strain KJ17, four for both BN06P03352 and GO13 strains.

Fig. 8

Conclusions

The current study reveals the comparative genomic analysis of *B*. *glycinifermentans* sp. with different *Bacillus* sp. To the best of our knowledge, in literature, there is no similar detailed study on *B*.

glycinifermentans sp. genomic features, functional annotation of the proteome, phylogenetic distinctness and evolutionary features. The results generated in this study provides valuable information to understand the phylogenetic relationship, functional annotation and genomic traits of this microorganism.

References

1. Schallmey, M., A. Singh, and O.P. Ward, *Developments in the use of Bacillus species for industrial production*. Canadian journal of microbiology, 2004. **50**(1): p. 1–17.
2. Rasko, D.A., et al., *Genomics of the Bacillus cereus group of organisms*. FEMS microbiology reviews, 2005. **29**(2): p. 303–29.
3. Liu, Y., et al., *Phylogenetic diversity of the Bacillus pumilus group and the marine ecotype revealed by multilocus sequence analysis*. PloS one, 2013. **8**(11): p. e80097.
4. Fitch, W.M., *Distinguishing homologous from analogous proteins*. Systematic zoology, 1970. **19**(2): p. 99–113.
5. Koonin, E.V., *Orthologs, paralogs, and evolutionary genomics*. Annual review of genetics, 2005. **39**: p. 309–38.
6. Gabaldon, T. and E.V. Koonin, *Functional and evolutionary implications of gene orthology*. Nature reviews. Genetics, 2013. **14**(5): p. 360–6.
7. Dessimoz, C., *Editorial: Orthology and applications*. Briefings in bioinformatics, 2011. **12**(5): p. 375–6.
8. Kim, S.-J., et al., *Bacillus glycinifermentans sp. nov., isolated from fermented soybean paste*. nt J Syst Evol Microbiol, 2015. **65**(10): p. 3586–3590.
9. Karim, A., et al., *Draft genome sequence of a novel Bacillus glycinifermentans strain having antifungal and antibacterial properties*. J. Glob. Antimicrob. Resist., 2019. **19**: p. 308–310.
10. Gurevich, A., et al., *QUAST: quality assessment tool for genome assemblies*. Bioinformatics, 2013. **29**(8): p. 1072–1075.
11. Seemann, T., *Prokka: rapid prokaryotic genome annotation*. Bioinformatics, 2014. **30**(14): p. 2068–9.
12. Lowe, T.M. and S.R. Eddy, *tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence*. Nucleic Acids Res, 1997. **25**(5): p. 955–964.
13. Lagesen, K., et al., *RNAmmer: consistent and rapid annotation of ribosomal RNA genes*. Nucleic Acids Res, 2007. **35**(9): p. 3100–3108.
14. Bankevich, A., et al., *SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing*. J. Comput. Biol., 2012. **19**(5): p. 455–477.
15. Sneath, P.H. and R.R. Sokal, *Numerical taxonomy. The principles and practice of numerical classification*. 1973.
16. Tamura, K., M. Nei, and S. Kumar, *Prospects for inferring very large phylogenies by using the neighbor-joining method.* Proc. Natl. Acad. Sci. U. S. A, 2004. **101**(30): p. 11030–11035.

17. Zuckerkandl, E. and L. Pauling, *Evolutionary divergence and convergence in proteins*, in *Evolving genes and proteins*, V. Bryson and H.J. Voge, Editors. 1965, Academic Press. p. 97–166.
18. Meier-Kolthoff, J.P., et al., *Genome sequence-based species delimitation with confidence intervals and improved distance functions*. BMC bioinformatics, 2013. **14**: p. 60.
19. Richter, M., Rosselló-Móra, Ramon, *Shifting the genomic gold standard for the prokaryotic species definition.* Proc. Natl. Acad. Sci. U. S. A, 2009. **106**(45): p. 19126–19131.
20. Kurtz, S., et al., *Versatile and open software for comparing large genomes*. Genome biology, 2004. **5**(2): p. R12.
21. Paradis, E., J. Claude, and K. Strimmer, *APE: Analyses of Phylogenetics and Evolution in R language*. Bioinformatics, 2004. **20**(2): p. 289–290.
22. Emms, D.M. and S. Kelly, *OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy*. Genome biology, 2015. **16**(1): p. 157.
23. Zhao, Y., et al., *PanGP: a tool for quickly analyzing bacterial pan-genome profile*. Bioinformatics (Oxford, England), 2014. **30**(9): p. 1297–1299.
24. Chaudhari, N.M., V.K. Gupta, and C. Dutta, *BPGA- an ultra-fast pan-genome analysis pipeline*. Sci. Rep., 2016. **6**: p. 24373.
25. Proost, S., et al., *i-ADHoRe 3.0–fast and sensitive detection of genomic homology in extremely large data sets.* Nucleic Acids Res, 2012. **40**(2): p. e11-e11.
26. Naquin, D., et al., *CIRCUS: a package for Circos display of structural genome variations from paired-end and mate-pair sequencing data*. BMC bioinformatics, 2014. **15**(1): p. 198.
27. Gevers, D., et al., *Opinion: Re-evaluating prokaryotic species*. Nature reviews. Microbiology, 2005. **3**(9): p. 733–9.
28. Altschul, S.F., et al., *Basic local alignment search tool*. Journal of molecular biology, 1990. **215**(3): p. 403–10.
29. Casimiro-Soriguer, C.S., A. Munoz-Merida, and A.J. Perez-Pulido, *Sma3s: A universal tool for easy functional annotation of proteomes and transcriptomes*. Proteomics, 2017. **17**(12).
30. Overbeek, R., et al., *Use of contiguity on the chromosome to predict functional coupling*. In silico biology, 1999. **1**(2): p. 93–108.
31. Sinha, A.U. and J. Meller, *Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms*. BMC bioinformatics, 2007. **8**: p. 82.
32. Vallenet, D., et al., *MaGe: a microbial genome annotation system supported by synteny results*. Nucleic Acids Res, 2006. **34**(1): p. 53–65.
33. Luis Carlos, G., et al., *Inside the Pan-genome - Methods and Software Overview*. Curr. Genomics. 2015. **16**(4): p. 245–252.
34. Medini, D., et al., *The microbial pan-genome*. Current opinion in genetics & development, 2005. **15**(6): p. 589–94.

35. Tettelin, H., et al., *Comparative genomics: the bacterial pan-genome*. Current opinion in microbiology, 2008. **11**(5): p. 472–7.

Figures

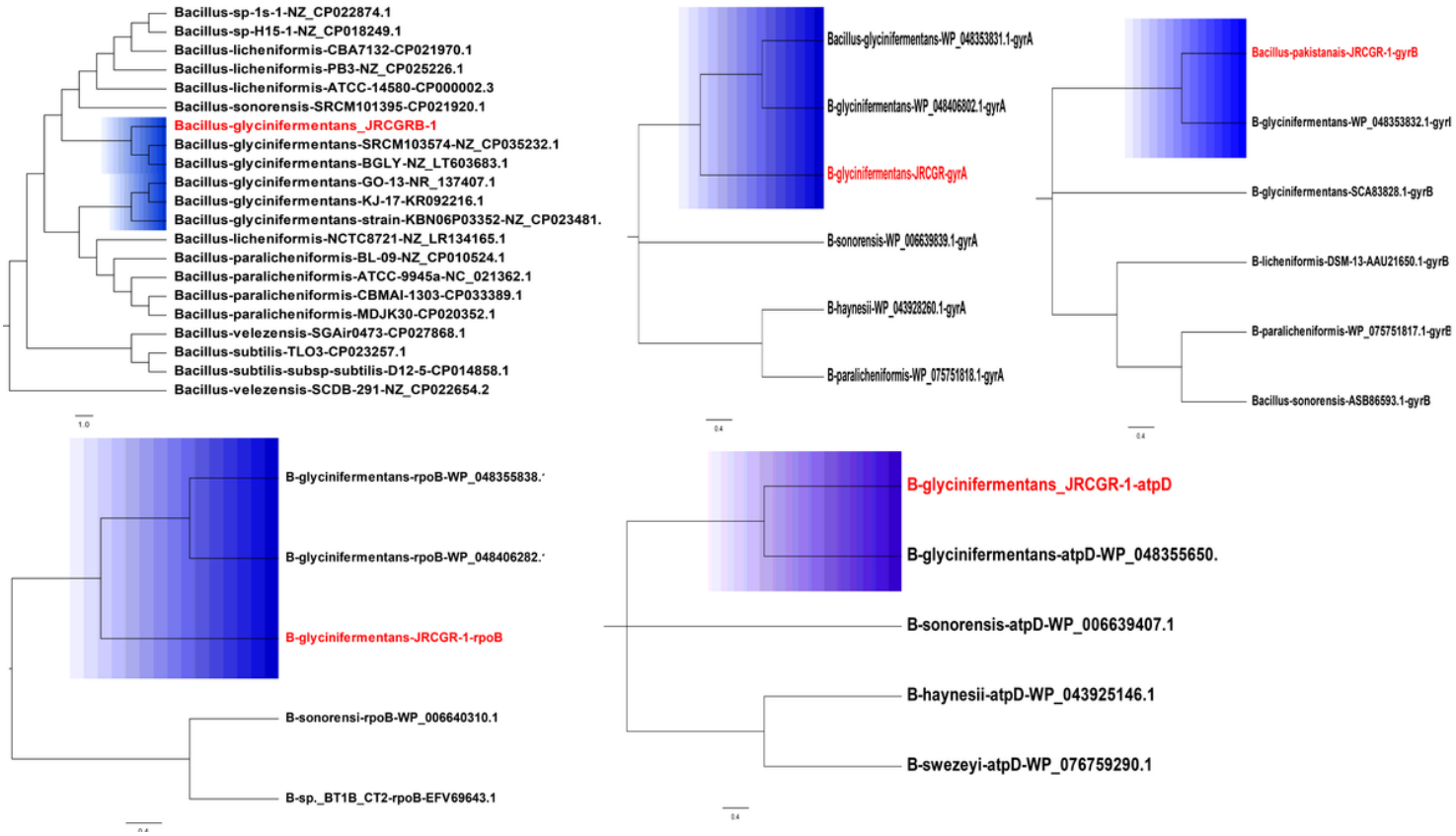

Figure 1

Phylogenetic tree base on 16S rRNA(A), gyrA(B), gyrB(C), rpoB(D) and aptD(E) genes.

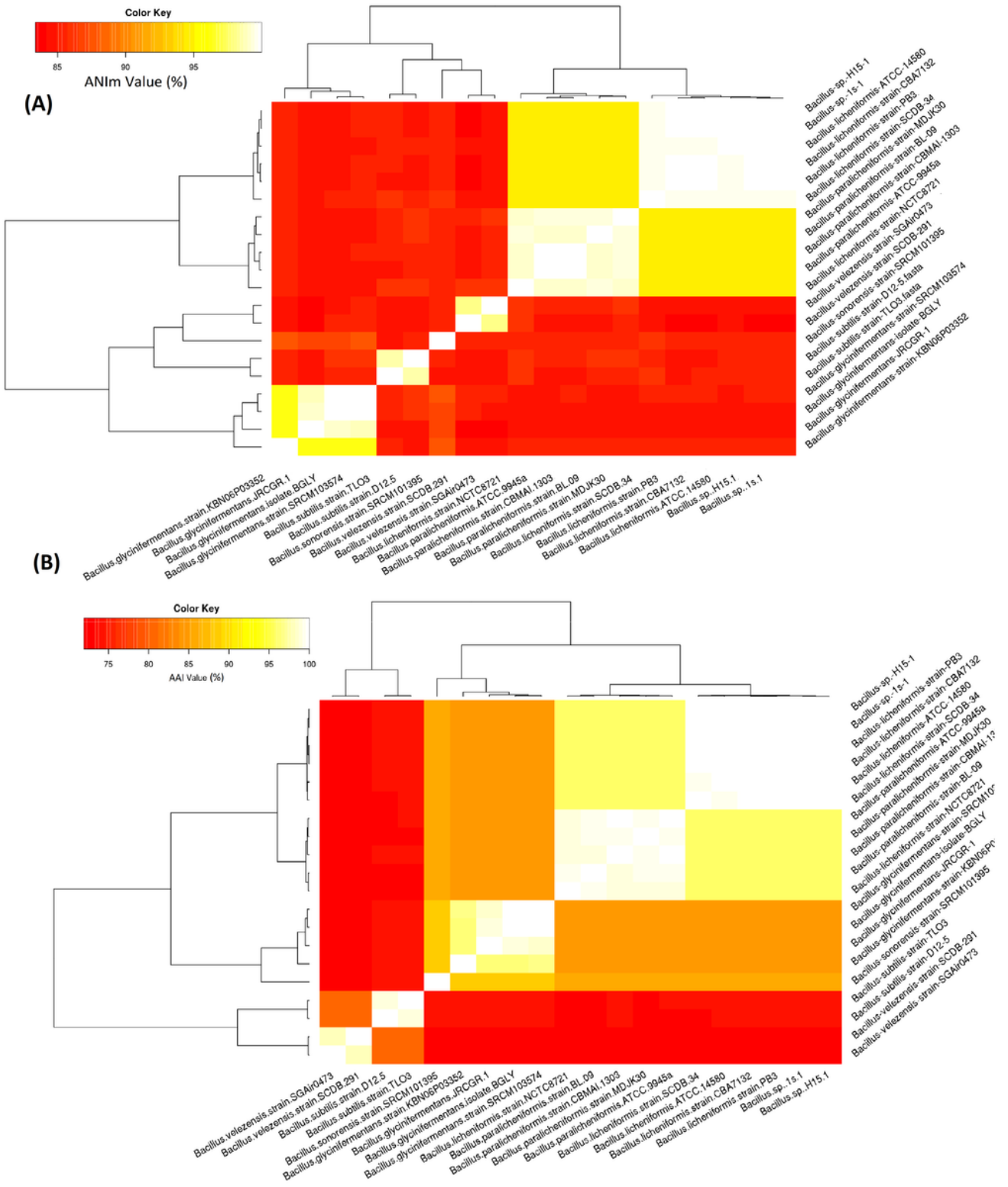

Figure 2

Species identification based on whole-genome sequencing: heatmap was derived from ANIm(A) and tetranucleotide signatures AAI (B).

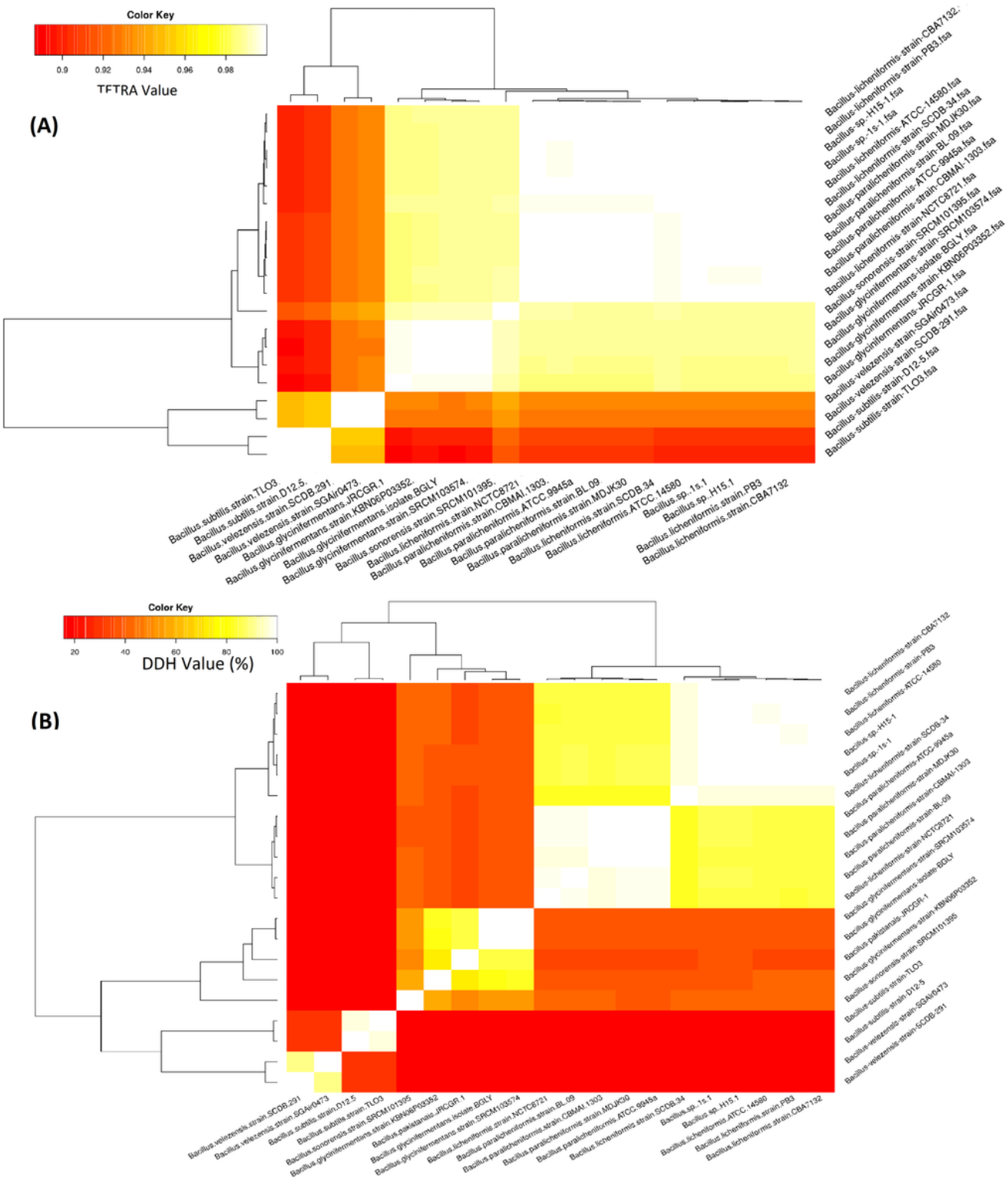

Figure 3

Species identification based on whole-genome sequencing: heatmap was derived (TETRA) (A) and in silico DDH (B) values.

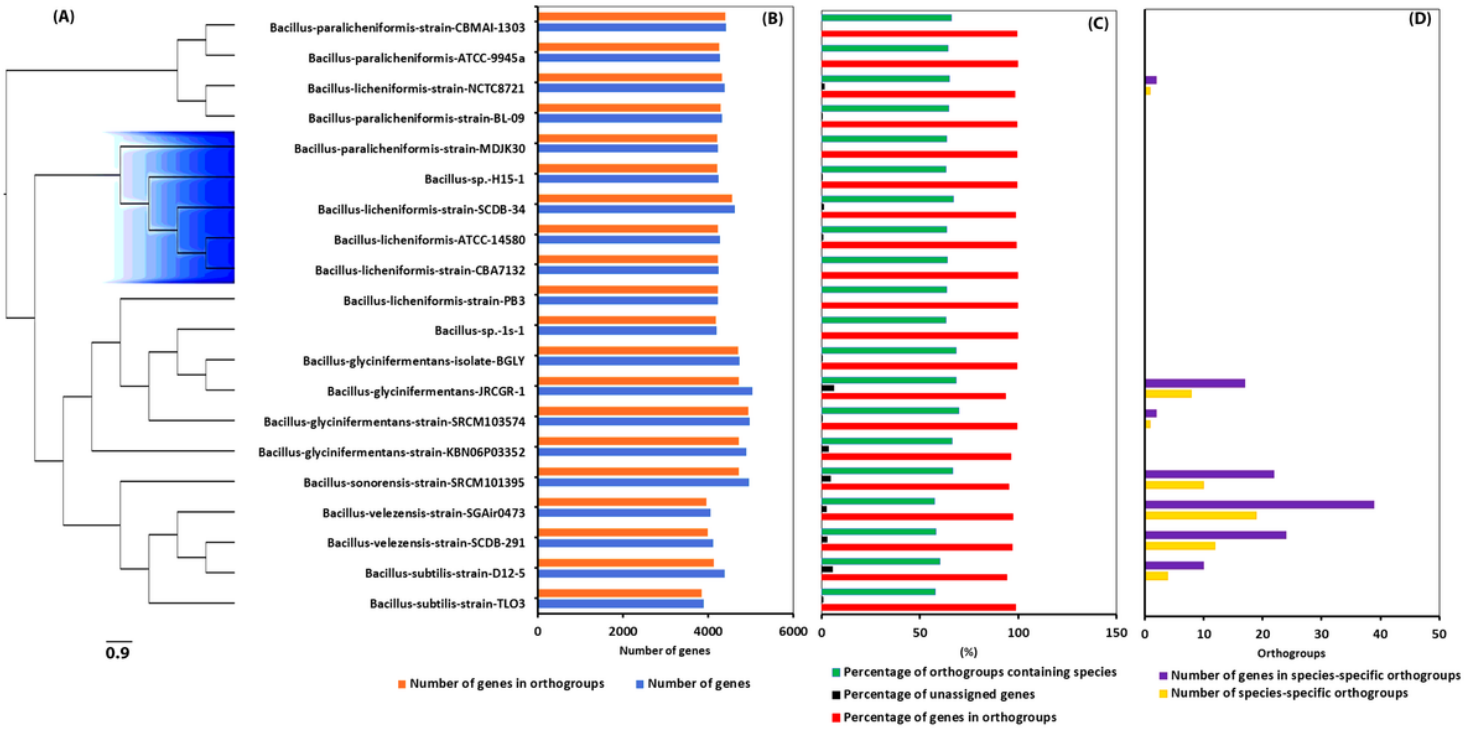

Figure 4

Summary of OrthoFinder analysis of a set of Bacillus sp. The species tree is inferred by STAG and rooted by STRID. B) Percentage of genes from each species assigned to orthogroups. C) The number of species-specific orthogroups. D) Percentage orthogroups containing species. E) Number of genes in orthogroups and number of genes in each Bacillus sp.

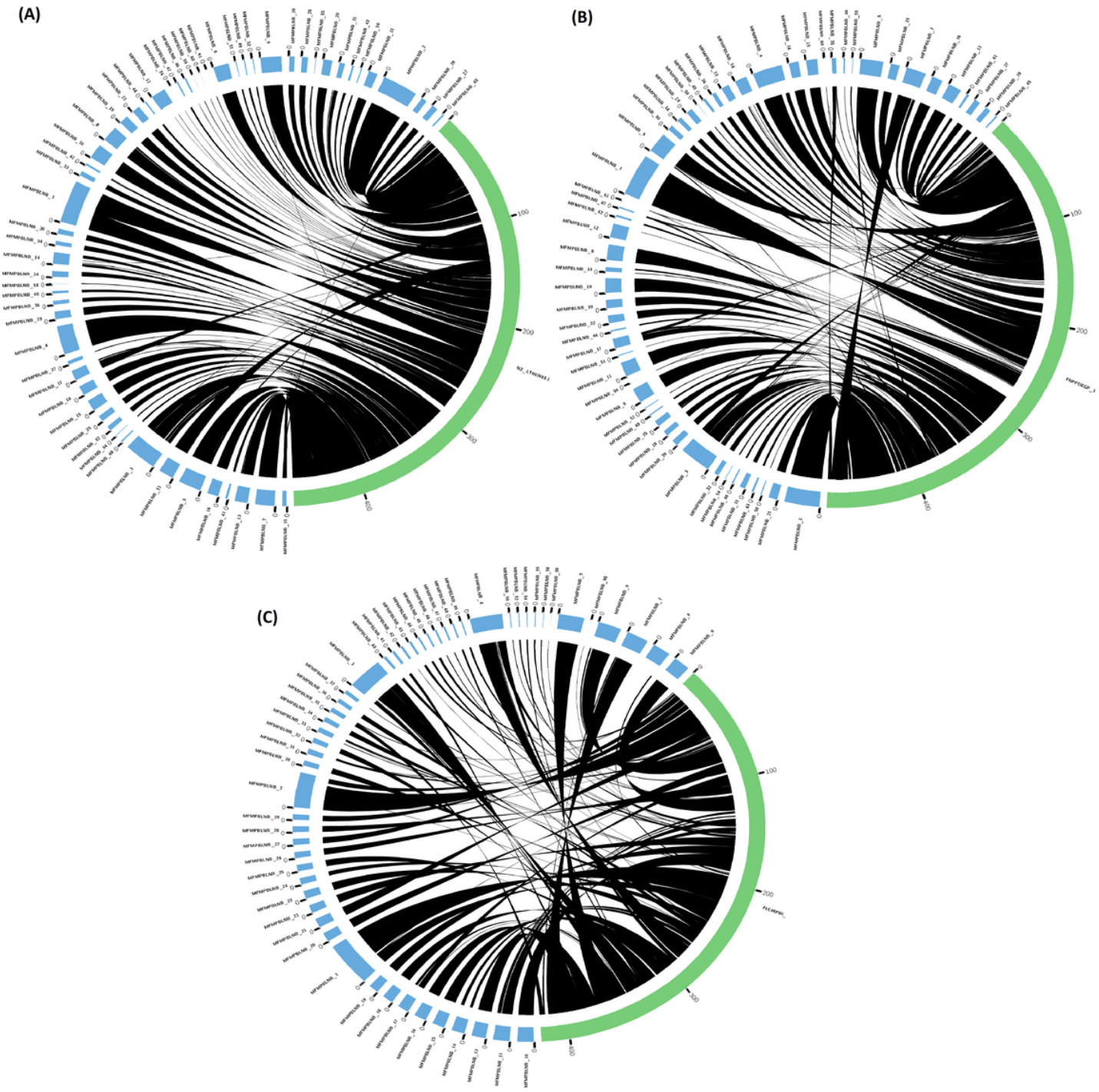

Figure 5

Synteny analyses for three primate genomes. Strain JRCGR-1 was used as reference genome and B. glycinifermentans BGLY (A), B. sonorensis strain SRCM101395 (B) and B. licheniformis strain ATCC 14580 (C) as a query genome

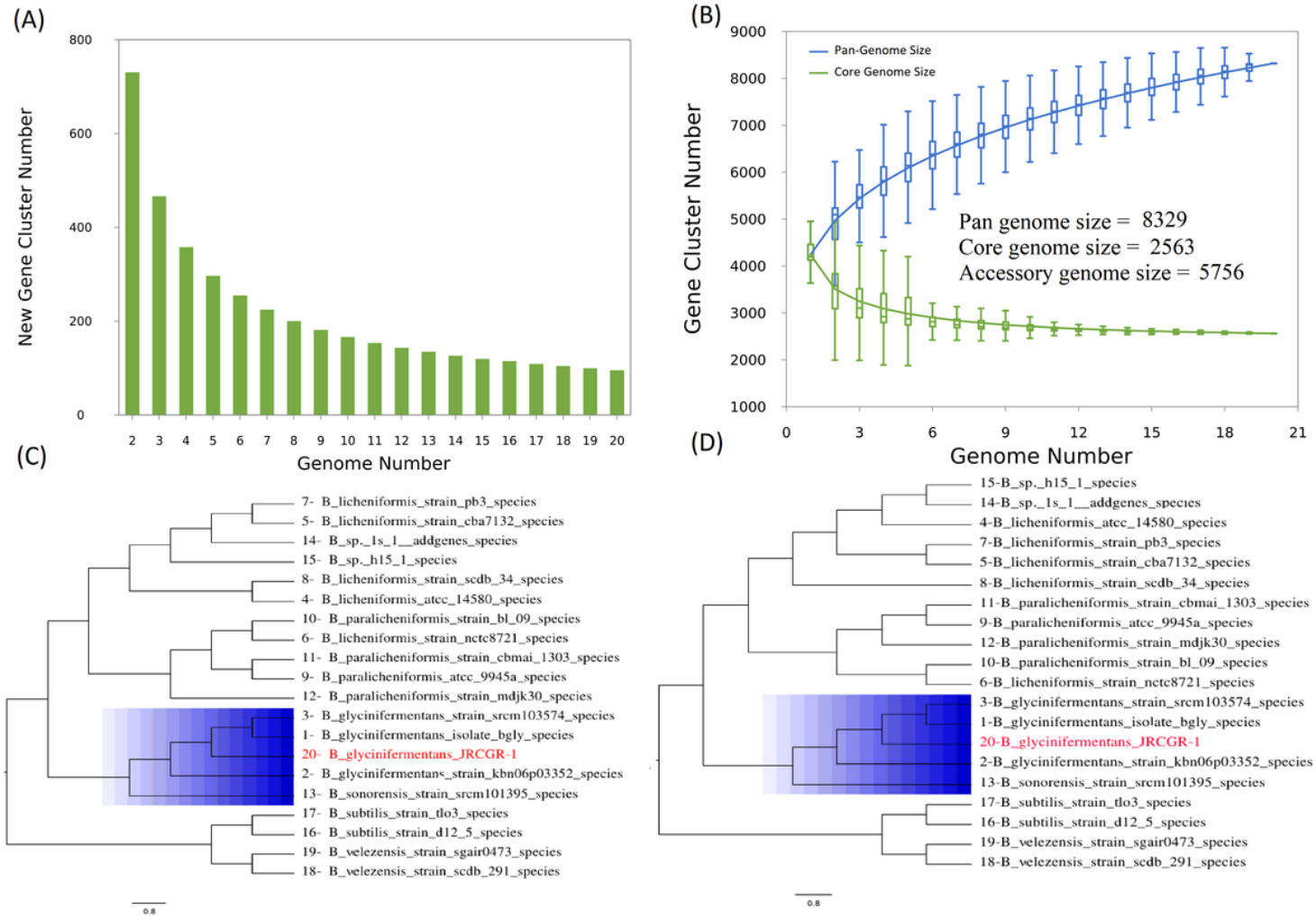

Figure 6

Pan-genome analysis. Curve for pan-genome and core genome of 20 Bacillus sp. (A). New gene family distribution after sequential addition of each genome to the analysis (B). Phylogenetic tree based on the core (C) and pan (D) genes.

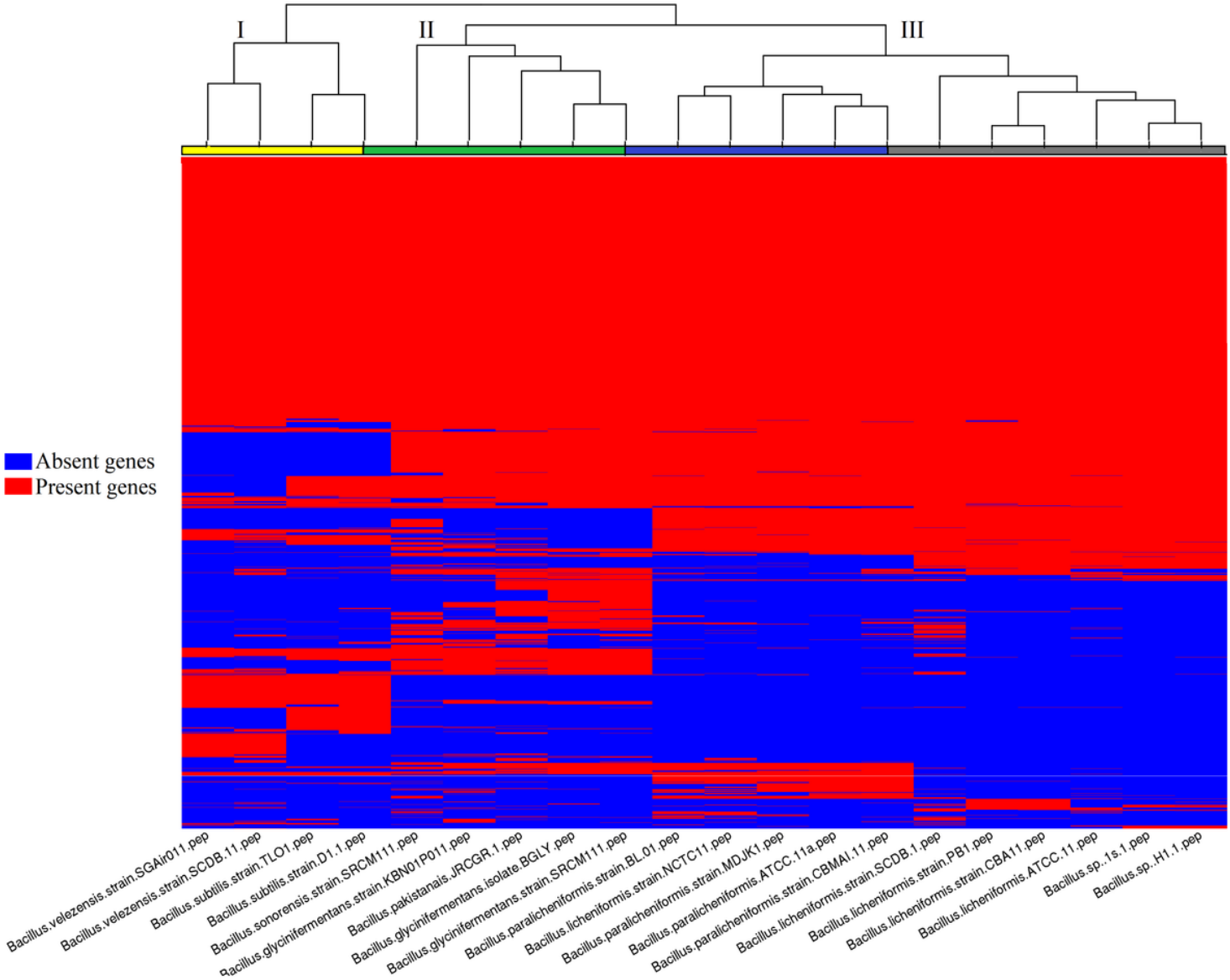

Figure 7

Heat map of total gene content comparisons and clustering. Gene presence is shown in blue and gene absence in red.

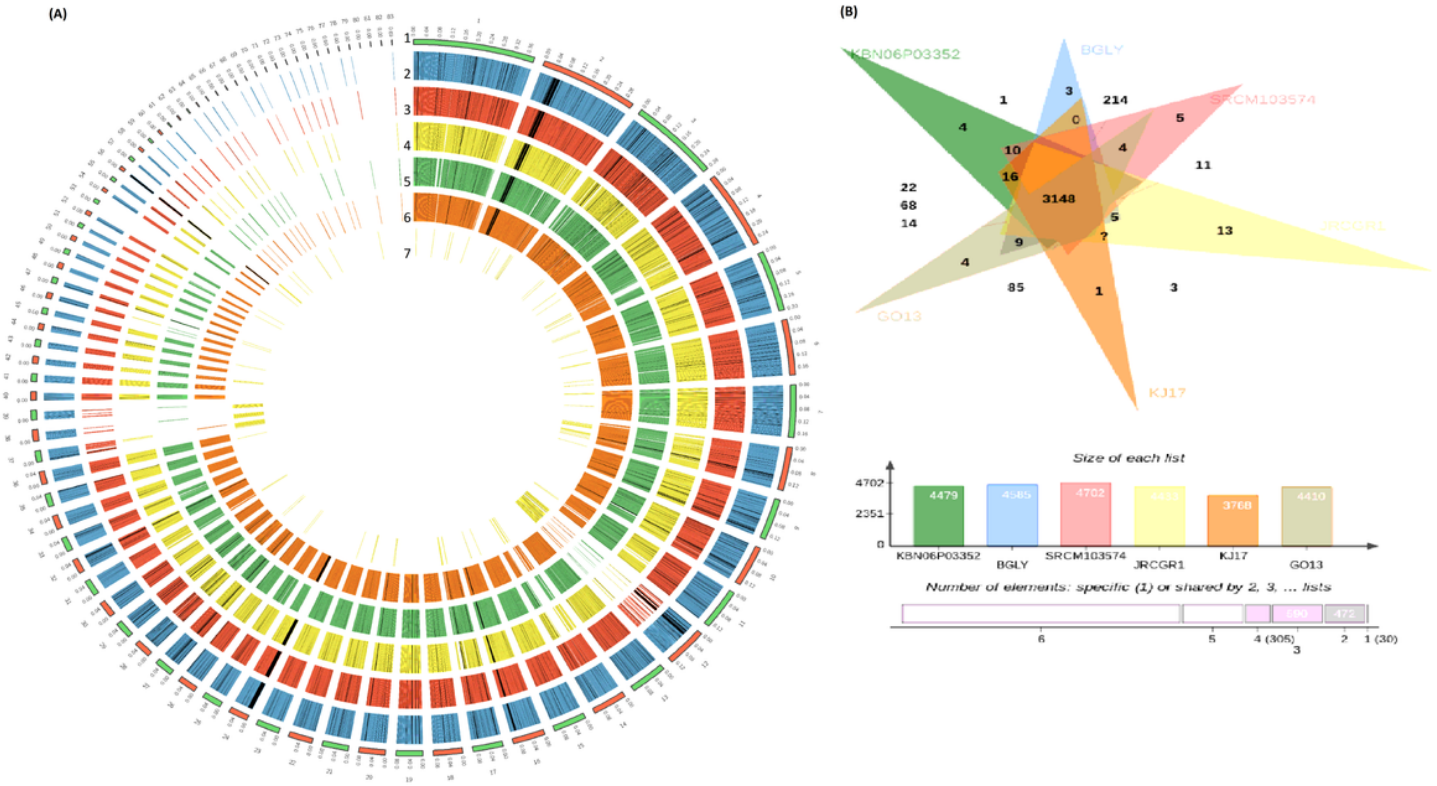

Figure 8

The graphical representation of orthologues genes for three B. glycinifermentans strains (A): from the outer circle inwards;83 scaffolds of strain JRCG-1 (1); genes (5074) of strain JRCG-1 (2); orthologues genes (4670) assigned to strain JRCG-1 (3); orthologues genes (4196) shared between JRCG-1 and BGLY strain (4); orthologues genes (4018) shared between JRCG-1 and KBN06PO3352 strain (5) orthologues genes (4274) shared between JRCG-1 and SRCM103574 strain (6). unassigned genes (375) of strain JRCG1 to any orthogroups (7). Venn diagram for the distribution of gene families (B).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1S.docx
- Table2S.docx
- Table3S.docx