Brazilian Journal of Microbiology (2022) 53:2051–2063
https://doi.org/10.1007/s42770-022-00815-0

ENVIRONMENTAL MICROBIOLOGY - RESEARCH PAPER

Phylogenomic characterization and pangenomic insights into the surfactin-producing bacteria *Bacillus subtilis* strain RI4914

Julie Kennya de Lima Ferreira[1] · Alessandro de Mello Varani[2] · Marcos Rogério Tótola[3] · Michelle Fernandes Almeida[3] · Dirceu de Sousa Melo[1] · Cristina Ferreira Silva e Batista[1] · Antonio Chalfun-Junior[1] · Kellen Kauanne Pimenta de Oliveira[1] · Luiz Fernando Wurdig Roesch[4] · Victor Satler Pylro[1]

Received: 29 April 2022 / Accepted: 22 August 2022 / Published online: 9 September 2022

Abstract

Bacillus subtilis is a versatile bacterial species able to produce surfactin, a lipopeptide biosurfactant. We carried out the phylogenomic characterization and pangenomic analyses using available *B. subtilis* complete genomes. Also, we report the whole genome of the biosurfactant-producing *B. subtilis* strain RI4914 that was isolated from effluent water from an oil exploration field. We applied a hybrid sequencing approach using both long- and short-read sequencing technologies to generate a highly accurate, single-chromosome genome. The pangenomics analysis of 153 complete genomes classified as *B. subtilis* retrieved from the NCBI shows an open pangenome composed of 28,511 accessory genes, which agrees with the high genetic plasticity of the species. Also, this analysis suggests that surfactin production is a common trait shared by members of this species since the *srfA* operon is highly conserved among the *B. subtilis* strains found in most of the assemblies available. Finally, increased surfactin production corroborates the higher *srfA*A gene expression in *B. subtilis* strain RI4914.

Keywords Biosurfactant · Nanopore sequencing · PGM IonTorrent, Whole genome sequencing · Hybrid assembly · GridIon

Introduction

Chemical surfactants cause environmental impacts mainly due to their toxicity and high environmental permanence [1]. Biosurfactants are amphiphilic molecules capable of reducing the surface and interfacial tension between compounds of different polarities [2]. Biosurfactants have become a promising alternative to their synthetic analogs due to their low toxicity, high biodegradability, and stability under extreme conditions, such as pH, salinity, and temperature [3–6]. These natural substances are used in several industrial sectors and are accepted and approved for commercialization. Many microbes able to produce biosurfactants have been isolated from soils or water [7], as they participate in carbon and nitrogen cycling in the ecosystem. These microbes produce diverse ecologically relevant biosurfactants with suitable characteristics for biotechnological applications.

Understanding the pathways and mechanisms involved in the production of biosurfactants is crucial to enhancing and extending the use of these organisms in industry. *Bacillus subtilis* strain RI4914 (a gram-positive and rod-shaped bacterium) is a good biosurfactant, solvent, and polymer producer [8]. This strain is a good candidate for application in microbial enhanced oil recovery (MEOR) since it recovered up to 88% of residual oil in laboratory conditions [8]. Surfactin is the most studied compound produced by *B. subtilis*. Surfactin participates in various applications, from antibiotic treatments (because of its

Responsible Editor: Luiz Henrique Rosa

✉ Victor Satler Pylro
victor.pylro@ufla.br

[1] Department of Biology, Federal University of Lavras - UFLA, Lavras, Minas Gerais, Brazil

[2] Departamento de Tecnologia, Faculdade de Ciências Agrárias E Veterinárias, Universidade Estadual Paulista (Unesp), Jaboticabal, Sao Paulo, Brazil

[3] Laboratório de Biotecnologia e Biodiversidade para o Meio Ambiente, Departamento de Microbiologia, Universidade Federal de Viçosa, Minas Gerais, Viçosa, Brazil

[4] Department of Microbiology and Cell Science, University of Florida, Gainesville, FL, USA

membrane permeabilizing properties) and cancer therapy to oil recovery processes. Biotechnological processes employing *Bacillus* sp. reported excellent Surfactin yields of more than one g L^{-1} [9–11]. Therefore, the large-scale industrial production of this compound may be viable through high cell density fed-batch cultivation.

Since the first genome sequencing strain of *Bacillus subtilis* in 1997, the species has become one of the most extensively studied and has the largest sequenced genomes. Four to five percent of the genome of *B. subtilis* is dedicated solely to the biosynthesis of natural products from numerous sources of carbon and nitrogen, such as animal fat, industrial effluents, burnt oil, whey, and water residues with a high starch content [12]. Many studies focus on pathogenic strains of the genus, such as *Bacillus anthracis* and *Bacillus cereus*. A pangenomic analysis of five different *Bacillus* species was performed to identify the properties of the core and strain-specific genes [13]. However, the analysis of core genes and specific genes of *Bacillus subtilis* strains is limited to few studies. Brito et al. [14] performed a pan genomic analyses of 42 *B. subtilis* genomes and suggested a process of gene acquisition driven by competence, where the maintenance of these genes depends on local and global fitness effects. As the cost of sequencing continues to decrease and the amount of *B. subtilis* complete genome sequence data generated grows, it is imperative to carry out more comprehensive analyses to characterize better the core and specific genes of *B. subtilis* strains [15, 16].

Here, to obtain more information about using the strain RI4914 for biosurfactant biosynthesis, we sequenced its genome using the GridION™ platform (Oxford Nanopore Technologies, UK) and the Ion Torrent PGM platform (Thermo Fisher). Subsequently, we perform high-quality hybrid genome assembly, annotation, and comparative genomics with other *B. subtilis* complete genomes available. This approach allowed us to design a new primer set for evaluating the expression of the *srfA*A gene, which codes for surfactin synthetase subunit 1, besides providing updated information regarding *B. subtillis* pangenomics.

Materials and methods

Bacillus subtilis strain RI4914 culture activation and cultivation for biosurfactant production under different growing conditions

Bacillus subtilis strain RI4914 was isolated from the production water of the Rio-Itaúnas formation, Conceição da Barra (ES, Brazil). It is available in the culture collection of the Laboratório de Biotecnologia e Biodiversidade para o Meio Ambiente (LBBMA), Department of Microbiology, Universidade Federal de Viçosa, UFV (Viçosa, MG, Brazil), as previously described [8].

The isolate was reactivated in Tryptone Soy Agar (TSA) at 30 °C for 24 h. A single colony was transferred to a tube containing 5 ml of Tryptone Soy Broth (TSB) in triplicate. The tubes were incubated in a shaker at 200 rpm and 30 °C for 18 h.

We transferred an aliquot of the pre-inoculum to 125 mL Erlenmeyer flasks containing 30 mL of mineral medium (13.9 g L^{-1} of KHP$_4$, 2.7 g L^{-1} of KH$_2$PO$_4$, 0.05 g L^{-1} of yeast extract, 4.24 g L^{-1} of NaNO$_3$, and 50 mL of micronutrient solution) supplemented with glucose (40 g L^{-1} — treatment 1) or synthetic glycerol (40 g L^{-1} — treatment 2) as a carbon source. We measured the microbial growth to obtain an optical density of 0.05 at 600 nm. The composition of the micronutrient solution was 0.5 g L^{-1} EDTA, 3.0 g L^{-1} MgSO4.7H$_2$O, 0.5 g L^{-1} MnSO$_4$.4H$_2$O, 1.0 g L^{-1} NaCl, 0. 1 g L^{-1} CaCl.2H$_2$O, 0.1 g L^{-1} CaCl$_2$.2H$_2$O, 0.1 g L^{-1} CaCl$_2$.6H2O, 0.1 g L^{-1} ZnSO$_4$.7H2O, 0.01 g L^{-1} CuSO$_4$. 5H$_2$O, 0.01 g L^{-1} Na$_2$MoO$_4$.2H2O, 0.01 g L^{-1} NaO$_4$Se, 0.01 g L^{-1} Na$_2$WO$_4$.2H$_2$O, and 0.02 g L^{-1} NiCl$_2$.6H$_2$O. This solution was autoclaved and added to sterilized culture media. The samples were incubated in triplicate in a shaker at 200 rpm at 30 °C for 48 h.

Oil spreading analysis

We used the oil spreading method to evaluate biosurfactant production following the procedure described by Youssef et al. (2004). The test was performed using two different carbon sources: glucose and synthetic glycerol. Briefly, we filled a petri dish (150 × 20 mm) with 100 ml of deionized water. Then, we deposited 30 µl of crude oil (Marlim Field, Rio de Janeiro, Brazil, API = 20.3) on the water surface to form an oil film. A 10-µl drop of the culture supernatant (without cells) was carefully deposited in the center of the oil film. The diameter of the clear zone formed in the oil film indicated the biosurfactant activity. The measurements were tested for normality with Shapiro–Wilk test, and the difference between treatments was evaluated by the *t*-test.

Extraction of nucleic acids from Bacillus subtilis RI4914

The DNA and RNA of the cultures under both treatments were coextracted using the RNA PowerSoil® Total RNA Isolation Kit (MoBio Laboratories, Carlsbad, CA) and RNA PowerSoil® DNA Elution Accessory Kit, following the manufacturer's instructions. We used Nanodrop Lite (Thermo Fisher Scientific, Waltham, MA, USA) to evaluate the purity of extracted nucleic acids. We used the Qubit 4.0 (Thermo Fisher Scientific, Waltham, MA USA) with the

Qubit™ dsDNA BR Assay Kit and Qubit™ RNA BR Assay Kit for nucleic acid quantification.

*srfA*A gene expression evaluation in Bacillus subtilis RI4914 under different growing conditions

We evaluate the expression of the *srfAA* gene, which codes for surfactin synthetase subunit 1 involved in protein biosynthesis in each treatment. Specific primers were designed using the list of genes recovered from *Bacillus subtilis* genomes as a reference. Briefly, we annotated the genomes obtained from GenBank using Prokka and retrieved the srfAA sequences. All sequences were aligned with Clustal X [17], and the NCBI Conserved Domain Search Tool (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) was used to verify the active domains. Then, BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) was used to verify the identity of the domains with the reference genome — *Bacillus subtilis* subsp. *subtilis* str. 168. The conserved domains were evaluated and avoided in the resulting PCR product. The selected sequences were used as input into Primer-BLAST (https://www.ncbi.nlm.nih.gov/tools/primer-blast/) to obtain the primer sequences. The parameters applied were PCR product size 80 (min) to 150 (max) and primer melting temperatures of 57 (min), 60 (opt), 63 (max), and 3 (Max Tmdiff). Each generated primer was tested via OligoAnalyzer™ (https://www.idtdna.com/calc/analyzer) to check for hairpin, self-dimer, and heterodimer formations. The parameters used were hairpin: DeltaG, positive is better; temperature: lower is better; and self-dimer: DeltaG, less than 10% of total DeltaG. Two pairs of oligonucleotide primers were obtained and synthesized, but one of them (srfAA-F: AGG CGGGGATCTTTGACA and srfAA-R: TGAAGCGGA ATCTCAATGC) showed better R2 and was selected for RT–qPCR analysis. The 16S rRNA gene was used as an endogenous control (16S-F: CCTACGGGAGGCAGCAG and 16S-R: ATTACCGCGGCTGCTGG).

RNA samples from both treatments were treated with the TURBO DNA-freeTM Kit (Thermo Fisher Scientific, Waltham, MA, USA) to eliminate residual DNA according to the manufacturer's recommendations for 5.0 µg of RNA. Furthermore, samples were quantified using a NanoVue® (GE Healthcare) spectrophotometer. cDNA synthesis was performed with the High-Capacity cDNA Reverse Transcription Kit (Thermo Fisher Scientific, Waltham, MA USA), following the manufacturer's recommendations for reactions without RNAse inhibitor.

RT–qPCR was performed once in a Rotor-Gene Q Real-Time PCR thermocycler (Qiagen) with an SYBR® Green detection system. The final volume of the reaction for each sample was 15 µL, including 7.5 µL of Master Mix SYBR Green (QuantiNova SYBR Green PCR Kit — Qiagen), 1.5 µL of cDNA in 1/5 dilution, 1.5 µL of each primer, and 3.0 µL RNase-free water. Three biological replicates of each treatment were used, with technical triplicates for each biological replicate, for each gene (target and reference). The cycling conditions used were 5 min at 95 °C for enzyme activation, followed by 40 cycles of 5 s at 95 °C and 10 s at 60 °C, with a melting temperature ramp-up to assess the specificity of the reaction between 55 and 95 °C, with an increase of 1 °C every 5 s. The efficiency of the target and reference gene primers was determined using a dilution curve. The relative expression was calculated using the method described by Pfaffl [18], with the 16S rDNA gene as a reference. For the gene expression statistical analysis, the expression rate and the confidence interval were calculated according to the method proposed by Steibel [19], which considers the linear mixed model given by the following equation:

$$yijkl = \mu + TGijk + \mathrm{II} + eijkl$$

where *yijklm* is the Cq (quantification cycle) obtained from the thermocycler software for the *k*th gene (reference or target) from the *m*th well, corresponding to the *l*th sample subject to the lth treatment; *TGijk* is the effect of the ith treatment (different growing media), in the expression of the gene *k* (reference or target). Statistical analysis was performed by the R software.

Genome sequencing, assembly, and annotation

Briefly, ~ 4 µg of unsheared genomic DNA was treated with the Rapid Sequencing Kit (SQK-RAD004; Oxford Nanopore Technologies, UK). The resulting library was sequenced on the GridION™ platform using a Spot-ON Mk1 flow cell (FLO-MIN 106 R9 version; Oxford Nanopore Technologies, UK) with an R9 Library Loading Bead Kit (EXP-LLB001; Oxford Nanopore Technologies, UK). The raw reads were acquired using MinKNOW software v3.5.6 in a 72-h experiment and base called using Albacore software v2.0.2. A subsample of the same DNA was sequenced using the Ion Torrent PGM platform (Thermo Fisher Scientific, Waltham, MA, USA). The template library was prepared with the Ion Plus Fragment Library Kit and clonally amplified in the One Touch 2 System with the Ion PGM™ Template Hi-Q OT2 400 Kit. The amplified library was sequenced using the Ion PGM™ Hi-Q Sequencing 400 Kit within a 316™ Chip v2. The nanopore sequencing (long reads) was performed at the Microbiology and Cell Science Department, University of Florida (UFL/USA). The Ion Torrent sequencing (short reads) was performed at the Laboratory of Microbial Ecology and Bioinformatics, Department of Biology, Federal University of Lavras (UFLA/Brazil).

The GridION™ long reads were de novo assembled using Canu v1.5 [20] following default parameters set for

Nanopore data. The low-quality GridION/Canu assembly was corrected (polished) with Racon [21] using high-quality PGM reads. For this purpose, PGM reads were mapped against the Canu assembly using the Burrows-Wheeler Alignment Tool (BWA) [22]. Then, Racon was run (parameters: -m 8 -x -6 -g -8 -w 500) using the Canu assembly as the "target sequence" to be corrected, the PGM reads as "correction sequences," and a SAM file (BWA output) containing overlaps between sequences and target sequences. A total of 3 rounds of polishing were performed. Genome completeness and contamination were estimated using CheckM [23] in lineage-specific mode. Digital DNA–DNA hybridization (dDDH) and average nucleotide identity based on BLAST (ANIb) between our genome and its reference, *Bacillus subtilis* subsp. *subtilis* str. *168*-NC_000964.3, was calculated using Genome-to-Genome Distance Calculator 2.1 (GGDC) [22, 24] and JspeciesWS [25], respectively. The *Bacillus subtilis* strain RI4914 genome annotation was performed with PATRIC version 3.5.23 [26] and with the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) [27].

Phylogenetic analyses of the *Bacillus subtilis* representatives

We retrieved all complete genome sequences of *B. subtilis* strains available in the GenBank database (https://www.ncbi.nlm.nih.gov) as of May 12, 2020, for a total of 153 genomes (Supplementary Table 1).

The complete *rpo*B gene (β subunit of the RNA polymerase) sequences were retrieved from the genomes under study and used for phylogenetic analysis. We used the type strain *B. amyloliquefaciens* DSM7 as the outgroup (accession number NC_014551.1). Sequence alignment was performed with muscle [28]. Bayesian inference (BA) was performed using the program MrBayes [29]. We chose the best nucleotide substitution model using the program MrModelTest2 (https://github.com/nylander/MrModeltest2). To infer the tree using BA, we ran two independent analyses with four chains each (one cold and three hot chains), starting with four random trees modified through 10,000,000 generations of MCMC. To construct the consensus tree, we checked the likelihood of the resulting topologies and burned out 25% of the trees (to retain those within the area with the highest likelihoods). The robustness of each tree node was obtained by the posterior probability, calculated by the frequency of each node in the consensus BA tree.

Pangenomic analyses

Pangenome analyses of all genomes were performed using the Roary software package v3.11.2 with default parameters [30]. Briefly, Prokka v1.14.6 [31] was used to annotate the assemblies with predicted genes. The annotated assemblies were then used as input for Roary to determine the core and accessory genes of all genomes.

Phylogenomics analyses

Multilocus sequence typing (MLST) was performed using the core gene set generated by Roary as the input. Two hundred and three core genes were concatenated and aligned with MAFTT [32]. The MLST analysis was performed with the maximum likelihood (ML) method using RAxML v8.2.10 [33] with default parameters, and the clade support estimates were calculated using rapid bootstrapping of 1000 pseudo-replicates.

Surfactin operon characterization in *B. subtilis*

The large ~ 27-kb *srfAABCD* operon is responsible for *B. subtilis* surfactin biosynthesis. We searched for this operon in the 153 annotated genomes used in this study to verify its predominance in *B. subtilis*. Additionally, we performed a phylogenetic analysis based on the complete operon with the maximum likelihood (ML) method using RAxML v8.2.10 [33] with default parameters. The clade support estimates were calculated using rapid bootstrapping of 1000 pseudo-replicates.

Accession numbers

The *B. subtilis* strain RI4914 Whole Genome Shotgun project is deposited at DDBJ/EMBL/GenBank under the accession number CP051306. The version described in this paper is the first version, CP051306.

Results and discussion

Surfactin production by *Bacillus subtilis* RI4914 and analysis of differential expression of the *srfAA* gene

Compared to those grown under the glycerol treatment, *Bacillus subtilis* RI4914 grown with glucose as the sole carbon source showed increased biosurfactant production, as measured by the diameter of the clear zone formed in the oil film (Figure S1; Fig. 1). On average, the oil spreading zone observed for *B. subtilis* grown in glucose was four times greater than the spreading zone observed for the same isolate grown in glycerol. Similar results were also found for isolates of other species of *Bacillus* [34]. Other *Bacillus* species are also reported as biosurfactant-producing bacteria. *Bacillus amyloliquefaciens* TSBSO 3.8 was able to produce Surfactin in a wide range of NaCl concentrations (0.5–7% w/v) and pH values (6 to 9), being a promising candidate

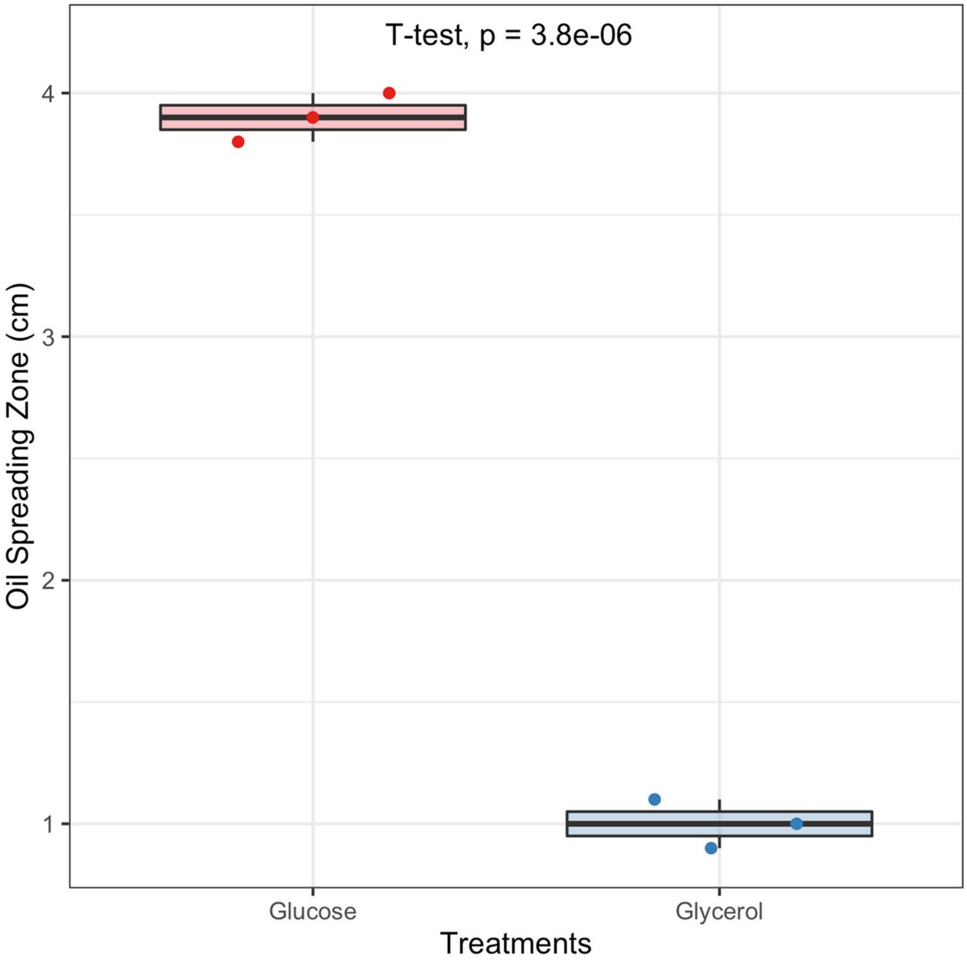

Fig. 1 Measurements of the oil spreading zone according to the different carbon sources. Boxes denote the interquartile range; the horizontal line inside the boxes represents the median

for application in MEOR [35]. Similarly, *Bacillus subtilis* SL has produced Surfactin at pH ranging from 5 to 12 and NaCl concentration of 10.0% (w/v), with a high potential for enhancing the oil recovery from low-permeability reservoirs [11].

The molecular basis that regulates these increases in biosurfactant production is still unknown. To further explore this finding, the expression of the *srfAA* gene was evaluated by RT–qPCR for bacteria grown under the same treatments with glucose or glycerol. Different pairs of primers published in the literature to assess the expression of surfactin genes were tested in this study; however, none showed sufficient amplification capacity and efficiency. To circumvent this problem, we designed and synthesized two pairs of primers targeting the *srfAA* gene and one pair targeting the 16S rRNA gene (endogenous control) for this study. The performance of these primers was evaluated (Table 1), and the best set was used for testing differential expression by RT–qPCR. Only one pair of primers produced a PCR product of the expected size. The primer efficiency for both pairs was equal to or greater than 100%. Still, the coefficient of correlation (R^2) obtained for the standard curve of the primer *2 srfAA* was low, indicating that the primers were not annealing to the target gene but rather to themselves, generating dimers (Table 1).

We observed an absolute increase in *srfA*A gene expression in the glucose treatment compared to the glycerol treatment (Fig. 2A). However, the difference was not significant at the 95% confidence level ($p = 0.0854$) according to the mixed linear statistical model proposed by Steibel [19] (Fig. 2B). This observation agrees with the results obtained through the oil spreading assay, suggesting that glucose induces an increased expression of the surfactin operon and, consequently, the production of surfactin by *Bacillus subtilis*.

B. subtilis strain RI4914 genome sequencing, assembly, and annotation

We obtained 3.48 M reads (20.69 Gb) from GridION™ sequencing, with 70 to 79,827 bp lengths. The reads were subsampled to 804,000, and those larger than 1000 bp were de novo assembled using Canu v1.5 [20] following default parameters set for Nanopore data. The genome was

Table 1 Performance of the primers targeting the *srfAA* and 16S rRNA (endogenous) genes

Primer set	Concentration (nmol)	R^2	Efficiency (%)
16S rRNA F: 5′-CCTACGGGAGGCAGCAG-3′ R: 5′-ATTACCGCGGCTGCTGG-3′	1	0.99408	86
*1srf*AA* F: 5′-ATGGCTTCATTCGTTCGGA-3′ R: 5′-GACGGTTCTCTTCAAGCC-3′	2	0.98867	112
2srfAA F: 5′-AGGCGGGGATCTTTGACA-3′ R: 5′-TGAAGCGGAATCTCAATGC-3′	2	0.02015	100

**srfAA* primer set selected for differential expression analysis

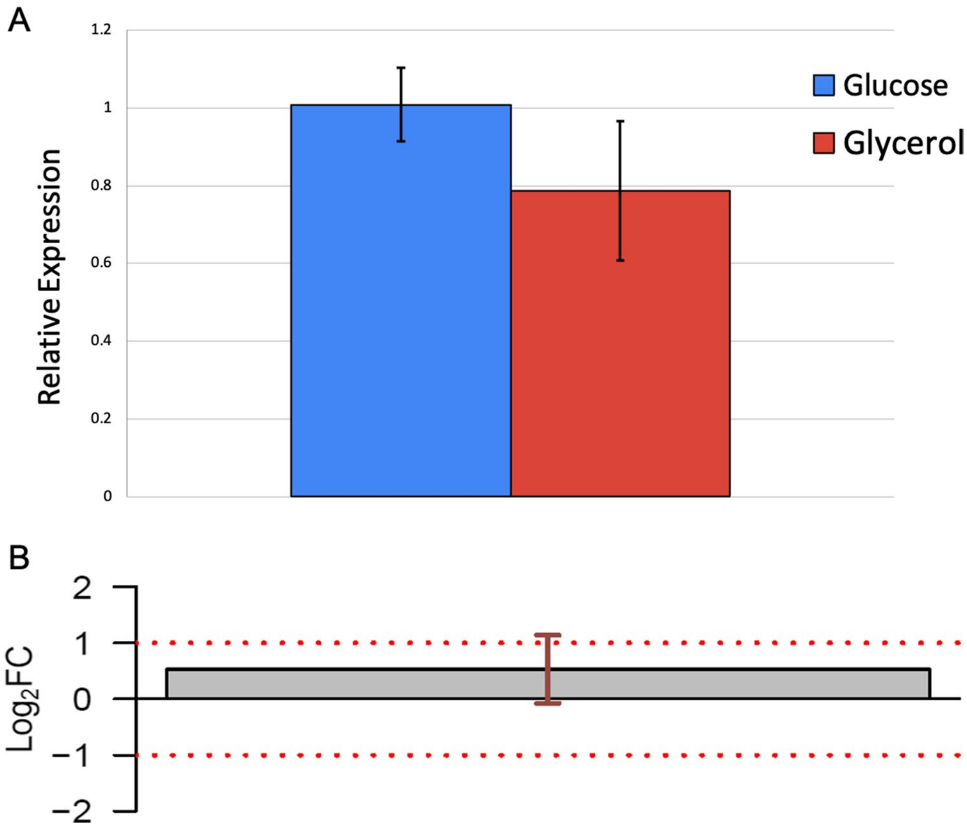

Fig. 2 **A** Relative expression of the *srf*AA gene in *Bacillus subtilis* RI4914 grown on glucose (blue) and glycerol (red). **B** Log2-fold change between the glucose and glycerol treatments

assembled into one complete circular chromosome. Similarly, 4.14 M high-quality reads (1.02 Gb, > Q20) were obtained from Ion Torrent PGM sequencing from 25 to 381 bp lengths. Both datasets were used for hybrid assembly.

The final genome consisted of a single contiguous circular chromosome (Fig. 3A). The assembly did not detect any plasmids. The genome size was 4,100,952 bp, with a G + C content of 43.48%. Genome completeness and contamination estimated by CheckM [23] were 98.84 and 1.10, respectively, indicating we obtained a complete with low contamination. The average nucleotide identity based on BLAST (ANIb) between our genome and the reference was 98.26. ANI scores of > 95% indicate that the sequences belong to the same species (Goris et al., 2007; Richter et al., 2015). The dDDH calculated using the genome-to-genome distance was 0.0157. This distance was transformed to values analogous to DDH using a generalized linear model (GLM) inferred from an empirical reference dataset comprising actual DDH values and genome sequences, which gave a value of 86.70% (84.1–88.9%) based on the formula (identities/HSP length) recommended by the program. The probabilities that DDH > 70% (i.e., same species) and DDH > 79% (i.e., same subspecies) were 94.53 and 59.84% (via logistic regression),

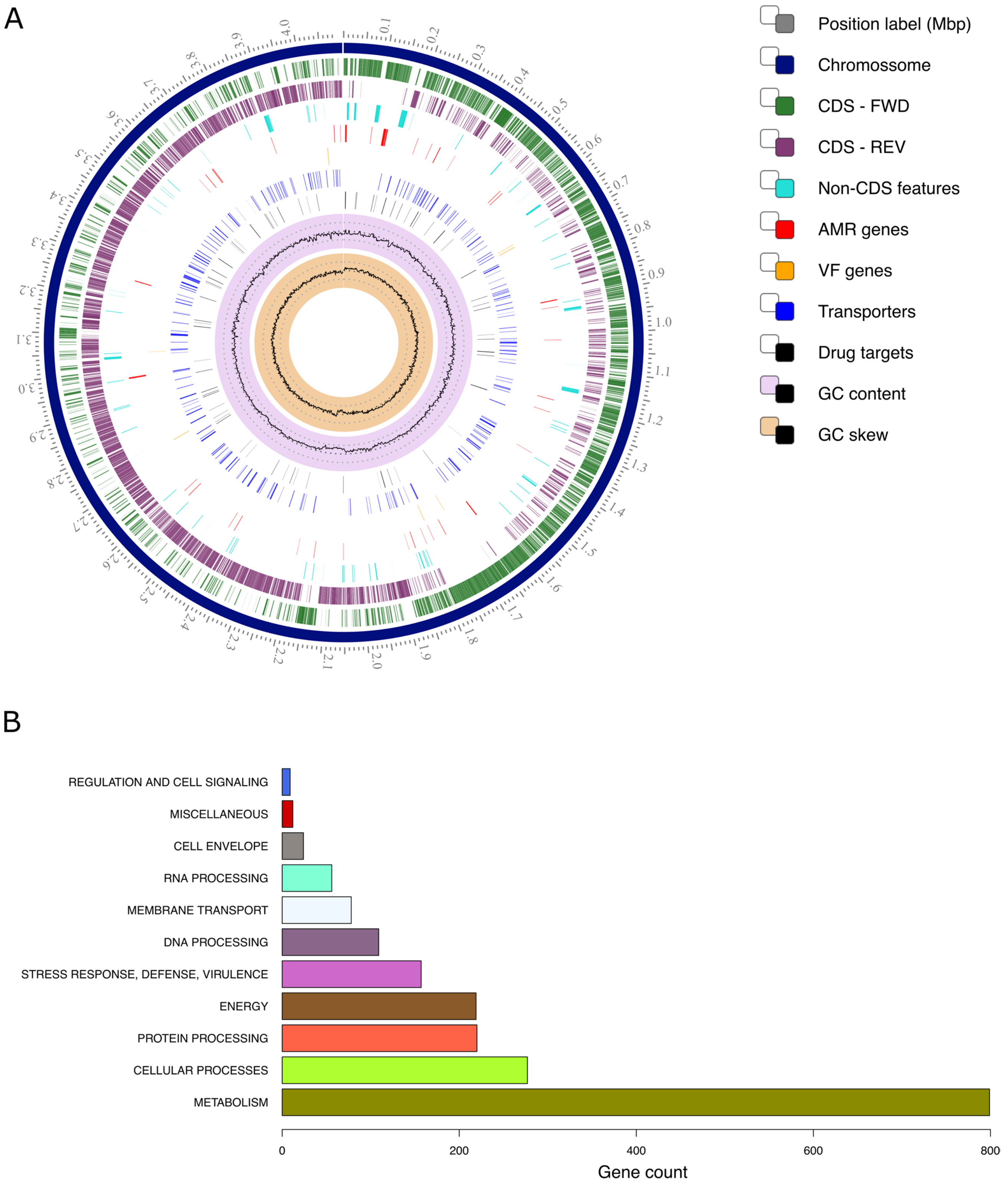

Fig. 3 **A** Circular graph view of the *Bacillus subtilis* strain RI4914 chromosome. **B** Class of subsystems classified by PATRIC for the *Bacillus subtilis* strain RI4914

respectively. The difference in the G + C content was 0.03 (same species).

A total of 4683 coding sequence sets (CDSs) and 116 predicted noncoding RNAs, including 86 tRNAs and 30 rRNAs (10 complete copies of the ribosomal operon), were identified by PATRIC. Functions were grouped into class subsystems by PATRIC (Fig. 3B). Most of the predicted genes were grouped in the "Metabolism" class, followed by "Cellular processes" and "Protein processing." Table S2 presents a complete list of subsystems associated with the *B. subtilis* strain RI4914. Finally, we searched for genes related to biosurfactant production. We found the entire *srfA* operon, including *srfAA*, *srfAB*, *srfAC*, and *srfAD*, which code for surfactin-producing proteins, and lipopeptides synthesized by nonribosomal peptide synthases (NRPSs). The annotation performed by PGAP is available at https://www.ncbi.nlm.nih.gov/nuccore/CP051306.

Phylogenetic and phylogenomic analysis of *B. subtilis*

Confirming the taxonomic classification of bacteria is extremely important, as the inclusion of misclassified bacteria may affect pangenomic studies [36]. To avoid misclassified bacteria in our analysis, we performed a phylogenetic and phylogenomic analysis before the pangenome analysis.

Figure 4 shows the Bayesian inference (BA) hypothesis for the *rpoB* gene. The phylogenetic reconstruction of microorganisms uses the *rpoB* gene [32, 33]. The *rpoB* gene is universal, conserved among bacteria and single copy, and recognized as a good marker gene for phylogenetic comparisons. In contrast, the 16S rRNA gene may have multiple copies, increasing sequence variations in a single organism. We were unable to recover the *rpoB* sequence from two of the analyzed genomes because it was not found in the annotated genome (NCBI: CP014858) or was frame shifted (NCBI: CP029052) (Table S3). These failures may be due to problems in annotation or poor quality of genome assembly. The *rpoB* gene is very conserved among all *B. subtilis* strains evaluated (Fig. 4). From the BA and similarity analysis, it was possible to confirm that strain RI4914 of *B. subtilis* is very close to the reference *B. subtilis* subsp. *subtle* str. 168 (NCBI: AL009126.3) isolated by Rasmusssem et al. [37], as they share 100% sequence similarity. The phylogeny proposed by the *rpoB* gene correlates with DNA-DNA hybridization (DDH) and ANI data. Sequence similarities for this gene greater than or equal to 97.7% significantly correlate with DDH values greater than 70% and ANI values greater than 94.3%, indicating these sequences are from the same species [38].

The low availability of *rpoB* gene sequences in public databases is still a limiting factor for phylogenetic reconstructions. The 16S rRNA gene has become widespread in genetic research due to the ease of sequencing the complete gene, approximately 1500 bp (3 times smaller than *rpoB*). However, Ki and colleagues [39] reported a

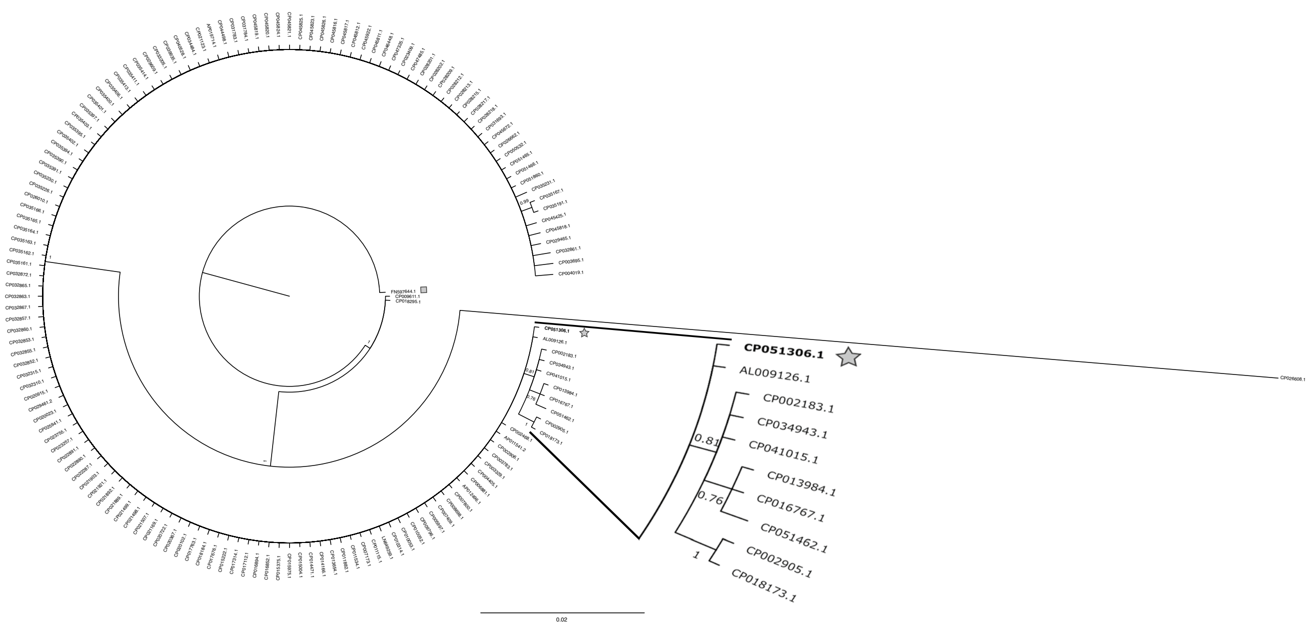

Fig. 4 Tree obtained by BA analysis from sequences of the *rpoB* gene. The values on the branches represent the posterior probability of the BA analysis. *Bacillus subtilis* RI4914 is marked with a star, and the type strain *Bacillus amyloliquefaciens* DSM7 used as the outgroup is marked with a square

resolution about 4.5 times greater than that of the 16S rRNA gene for the *rpoB* in phylogenetic analyses in *Bacillus,* providing an improved identification scheme for this group.

As DNA sequencing technologies have advanced, the number of complete genomes has grown significantly (e.g., NCBI: https://www.ncbi.nlm.nih.gov/genbank/statistics/); this allows the recovery of genes of interest, such as *rpoB*, and may improve our capacity to reconstruct microbial phylogenies.

To confirm the results obtained by the phylogenetic analysis, we also performed a phylogenomic analysis with 263 core genes from the complete genome sequences of *B. subtilis* strains retrieved from the GenBank database. The 263 core genes were concatenated, aligned, and used for phylogenomic analysis (MLST) (Fig. 5) through the maximum likelihood (ML) method. *B. subtilis* RI4914 clustered with strain JAAA, isolated from estuarine water [40], and strain SRCM103581, isolated from fermented soy food (https://www.ncbi.nlm.nih.gov/nuccore/CP035403.1). Core genes were more discriminatory for cluster analysis than single housekeeping genes, such as the *rpoB* gene.

Altogether, the results confirmed the taxonomic classification of the bacteria used in our study, allowing us to further explore the pangenome based on all 153 complete genome sequences of *B. subtilis* strains available in the GenBank database (https://www.ncbi.nlm.nih.gov) as of 12 May 2020.

Pangenome: comparing the genome of the *B. subtilis* strain RI4914 with similar genomes

In this type of analysis, core genes, accessory genes, and single genes (specific to each strain) are usually found and organized into soft-core, shell, and cloud genes [41]. The soft-core genes represent conserved genes in at least 95–99% of all analyzed genomes. Shell genes are those present in 15–95% of genomes, and cloud genes are those in 0–15% of genomes. The core and soft-core genes represent a pool of highly conserved genes that can provide information about the evolutionary history of the studied species. The cloud genes are those specific or unique to each strain present in the pangenomic analysis and are shared with at most two strains. The remaining genes are moderately conserved, identified as shell genes that, together with cloud genes, represent the subset of the accessory genome, which is responsible for reflecting the lifestyle, adaptation, and evolutionary history of the strains under study [41, 42].

The pangenome analyses of the 153 *B. subtilis* genomes were performed using Roary. We identified a core genome (genes shared by at least 151 genomes) composed of 263 genes and an accessory genome consisting of 28,511 genes, divided into subgroups of soft-core, shell, and cloud genes (Fig. 6A). The *B. subtilis* pangenome was classified as open since the number of unique genes in the pangenome continued to increase as a new genome was added to the analysis (Fig. 6B).

Bacillus subtilis is a gram-positive bacterium widely distributed and commonly described in soil and water [7]. Our results agree with the *Bacillus subtilis* high genetic plasticity and good ability to incorporate new genetic information into the genome due to intra- or interspecific horizontal gene transfer. Thus, providing the species with more significant genetic variability even under drift and natural selection processes [43]. This ability is a striking feature in prokaryotic species, such as *B. subtilis*, that have the potential to colonize different niches and increase their adaptation to a variety of environments [36]. Such adaptations can be related to the functionality of accessory genes, which can be neutral or beneficial. However, under some recombination processes and mutations, these adaptations may result in new traits, such as antibiotic resistance, survival mechanisms in stressful situations, or even increased virulence [44].

A similar pangenomic profile was reported for 61 genomes of *B. subtilis* strains isolated from fermented foods sold in East Asian countries (29 strains isolated from China, three from Japan, 24 from South Korea, and five from their study) [45]. The authors obtained 2098 core genes among the 61 strains and 3275 among the five strains isolated in their research and 6061 unique genes in an open pangenome.

Surfactin operon characterization in *B. subtilis*

We searched for the complete *srfA* operon, including *srfAA* (Surfactin synthase subunit 1), *srfAB* (Surfactin synthase subunit 2), *srfAC* (Surfactin synthase subunit 3), and *srfAD* (surfactin synthase thioesterase subunit) (Fig. 7A), encoding for surfactin-producing proteins, which are lipopeptides synthesized by nonribosomal peptide synthases (NRPS). Of the 153 genomes, 146 presented the annotated *srfA* operon (Table S4), showing that this function is a common trait among the *B. subtilis* representatives. Figure 7B shows the BA hypothesis for this operon. The *srfA* operon of the *B. subtilis* RI4914 strain was more closely related to that in *B. subtilis* JAAA (NCBI: CP045425), isolated from estuarine water. We observed that the *srfA* operon sequence was not highly conserved among the analyzed genomes, but further study is needed to verify whether these variations are natural or result from misassembly issues since we recovered genomes already assembled and not raw SRA data from the NCBI database.

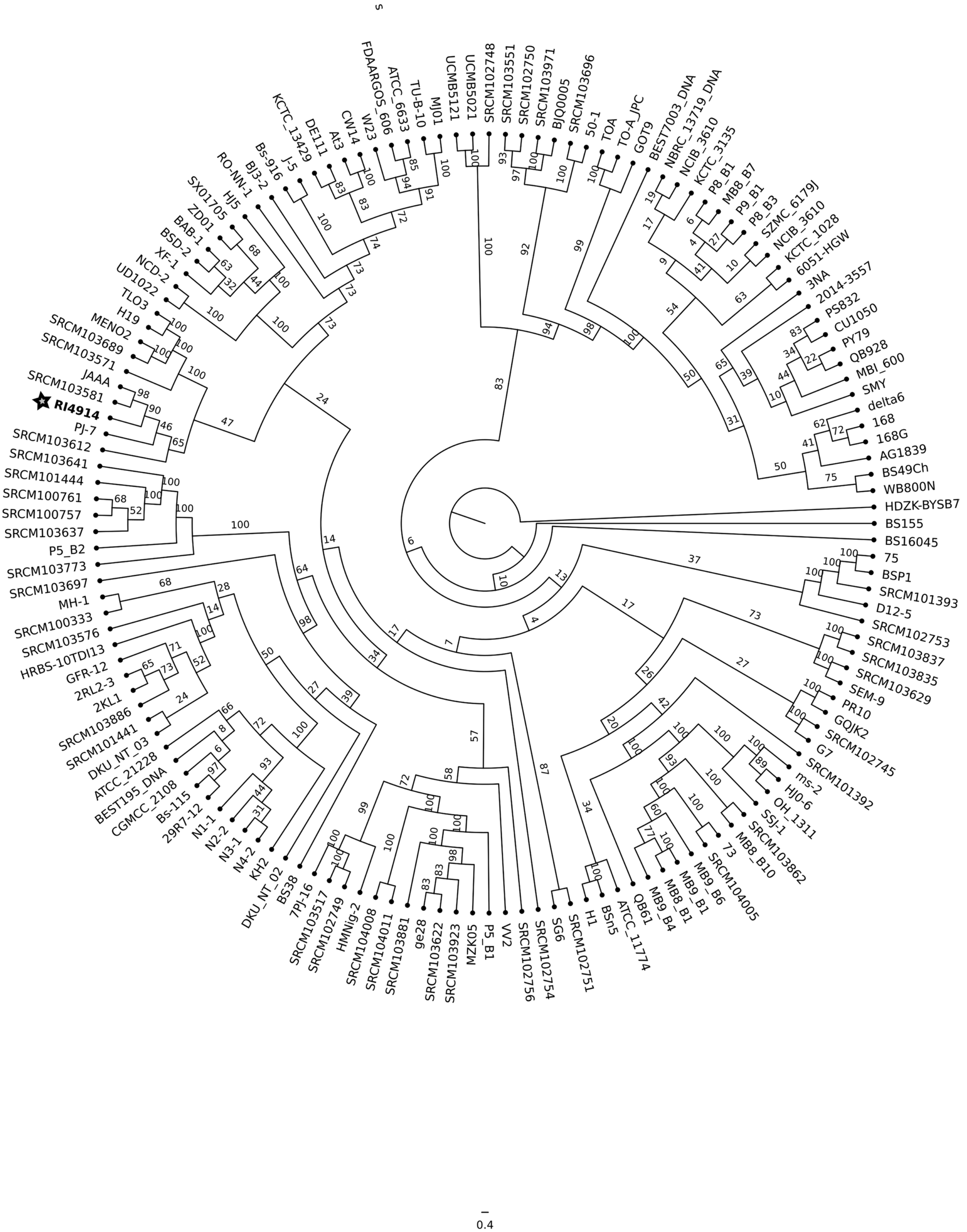

Fig. 5 Tree obtained by ML using the concatenated 263 core genes (MLST). The values on the branches represent the bootstrap values of 1000 pseudo-replicates. The *Bacillus subtilis* strain RI4914 is marked with a star

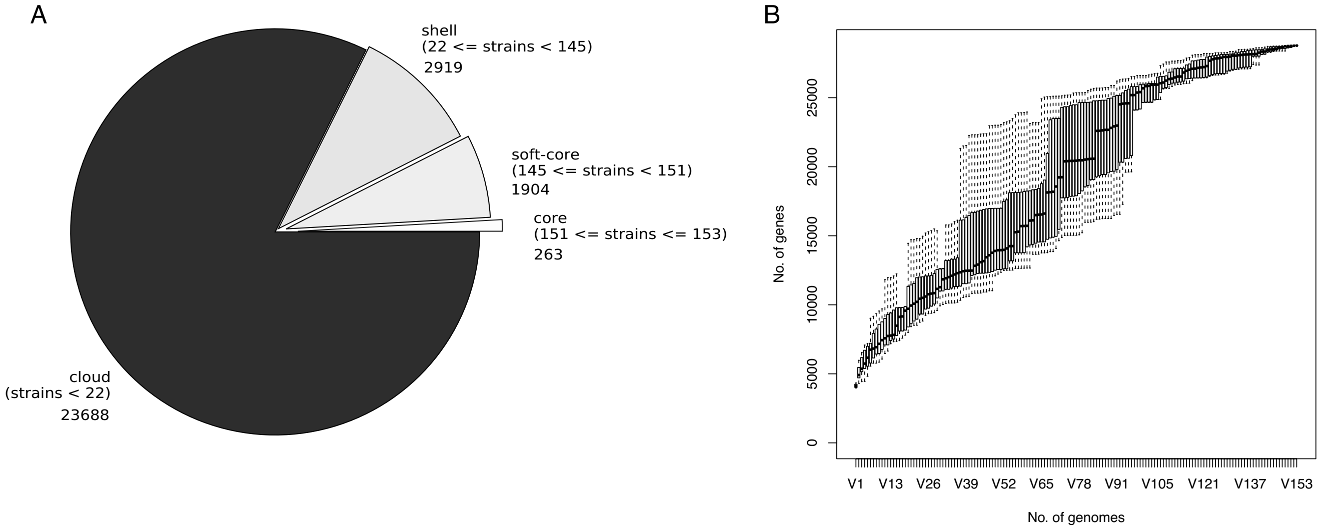

Fig. 6 **A** *Bacillus subtilis* pangenome showing the number of core, soft-core, shell, and cloud genes. **B** Open pangenome of *Bacillus subtilis*. The number of genes in the pangenome increases with the number of newly sequenced strains

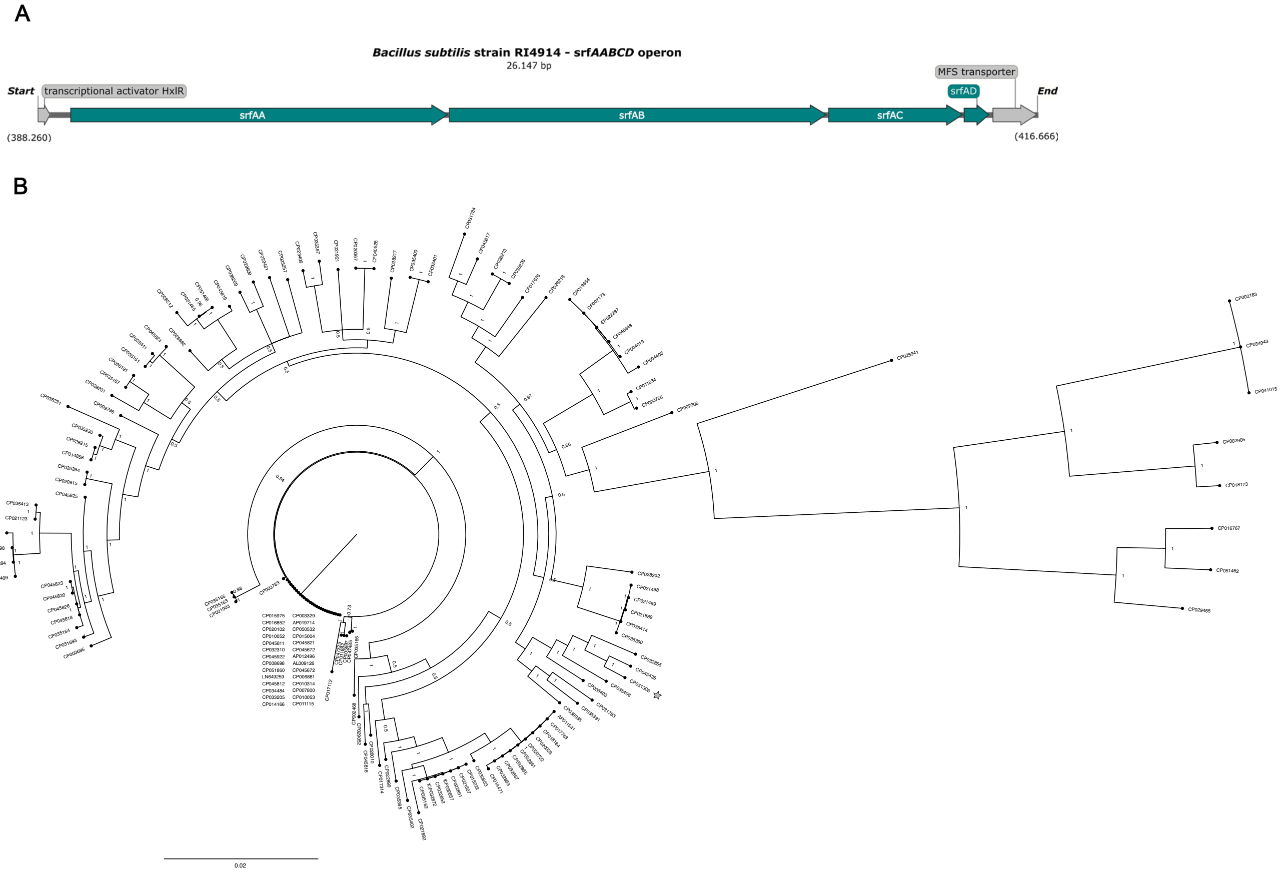
A
Bacillus subtilis strain RI4914 - srfAABCD operon
26.147 bp
Start
transcriptional activator HxlR
MFS transporter
srfAD
End
srfAA
srfAB
srfAC
(388.260)
(416.666)
B
0.02

Fig. 7 **A** Map of the *srfaA* operon in the *Bacillus subtilis* RI4914 strain. **B** Tree obtained by BA analysis from the concatenated core genes. The values on the branches represent the posterior probabilities of the BA analysis. *Bacillus subtilis* RI4914 is marked with a star

The genomic characterization of functional genes like those in the *srfA* operon may help us optimize genetic engineering approaches aiming for obtaining mutants with increased Surfactin production capacities. Also, it has allowed us to design new primers for analyzing the *srfA*A gene expression since those already available in the literature have failed in efficiency and/or the coefficient of correlation (R^2) obtained for the standard curve. Finally, our study provides updated information regarding *B. subtillis* pangenomics as the amount of complete genomes generated has grown over the last few years.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42770-022-00815-0.

Funding This work was supported by the National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico/CNPq—project grant number 404651/2018–6). AMV received a research fellowship from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (#303061/2019–7). JKLF also received a research fellowship from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (#133550/2019–2). This research was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior — Brasil (CAPES) — Finance Code 001. This work was also supported by the Brazilian Microbiome Project (http://www.brmicrobiome.org).

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Van Hamme JD, Singh A, Ward OP (2006) Physiological aspects. Biotechnol Adv 24:604–620. https://doi.org/10.1016/j.biotechadv.2006.08.001
2. Abbasi H, Hamedi MM, Lotfabad TB, Zahiri HS, Sharafi H, Masoomi F et al (2012) Biosurfactant-producing bacterium, Pseudomonas aeruginosa MA01 isolated from spoiled apples: physicochemical and structural characteristics of isolated biosurfactant. J Biosci Bioeng 113:211–219
3. Desai JD, Banat IM (1997) Microbial production of surfactants and their commercial potential. Microbiol Mol Biol Rev 61:47–64
4. Lima TMS, Procópio LC, Brandão FD, Leão BA, Tótola MR, Borges AC (2011) Evaluation of bacterial surfactant toxicity towards petroleum degrading microorganisms. Bioresour Technol 102:2957–2964
5. Lima TMS, Procópio LC, Brandão FD, Carvalho AMX, Tótola MR, Borges AC (2011) Biodegradability of bacterial surfactants. Biodegradation 22:585–592. https://doi.org/10.1007/s10532-010-9431-3
6. Maia M, Capão A, Procópio L (2019) Biosurfactant produced by oil-degradingPseudomonas putidaAM-b1 strain with potential for microbial enhanced oil recovery. Bioremediat J 23:302–310
7. Harwood CR, Mouillon J-M, Pohl S, Arnau J (2018) Secondary metabolite production and the safety of industrially important members of the Bacillus subtilis group. FEMS Microbiol Rev 42:721–738
8. Fernandes PL, Rodrigues EM, Paiva FR, Ayupe BAL, McInerney MJ, Tótola MR. (2016) Biosurfactant, solvents and polymer production by Bacillus subtilis RI4914 and their application for enhanced oil recovery. Fuel 180:551–557. https://doi.org/10.1016/j.fuel.2016.04.080
9. Willenbacher J, Yeremchuk W, Mohr T, Syldatk C, Hausmann R (2015) Enhancement of Surfactin yield by improving the medium composition and fermentation process. AMB Express 5:145
10. Zhi Y, Wu Q, Xu Y (2017) Genome and transcriptome analysis of surfactin biosynthesis in Bacillus amyloliquefaciens MT45. Sci Rep 7:40976
11. Wu B, Xiu J, Yu L, Huang L, Yi L, Ma Y (2022) Biosurfactant production by Bacillus subtilis SL and its potential for enhanced oil recovery in low permeability reservoirs. Sci Rep 12:7785
12. Madsen JK, Pihl R, Møller AH, Madsen AT, Otzen DE and Andersen KK (2015) The anionic biosurfactant rhamnolipid does not denature industrial enzymes. Front Microbiol 6:292. https://doi.org/10.3389/fmicb.2015.00292
13. Kim Y, Koh I, Young Lim M, Chung W-H, Rho M (2017) Pan-genome analysis of Bacillus for microbiome profiling. Sci Rep 7:10984
14. Brito PH, Chevreux B, Serra CR, Schyns G, Henriques AO, Pereira-Leal JB (2018) Genetic competence drives genome diversity in Bacillus subtilis. Genome Biol Evol 10:108–124
15. Borriss R, Danchin A, Harwood CR, Médigue C, Rocha EPC, Sekowska A, et al. (2018) Bacillus subtilis,the model Gram-positive bacterium: 20 years of annotation refinement. Microb Biotechnol 11:3–17. https://doi.org/10.1111/1751-7915.13043
16. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V et al (1997) The complete genome sequence of the gram-positive bacterium Bacillus subtilis. Nature 390:249–256
17. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948. https://doi.org/10.1093/bioinformatics/btm404
18. Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res 29:e45
19. Steibel JP, Poletto R, Coussens PM, Rosa GJM (2009) A powerful and flexible linear mixed model framework for the analysis of relative quantification RT-PCR data. Genomics 94:146–152
20. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM (2017) Canu: scalable and accurate long-read assembly via adaptive -mer weighting and repeat separation. Genome Res 27:722–736
21. Vaser R, Sović I, Nagarajan N, Šikić M (2017) Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res 27:737–746
22. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760
23. 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
24. Meier-Kolthoff JP, Carbasse JS, Peinado-Olarte RL, Göker M (2021) TYGS and LPSN: a database tandem for fast and reliable genome-based classification and nomenclature of prokaryotes. Nucleic Acids Res. https://doi.org/10.1093/nar/gkab902
25. Richter M, Rosselló-Móra R, Oliver Glöckner F, Peplies J (2016) JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. Bioinformatics 32:929–931
26. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL et al (2014) PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res 42:D581–D591
27. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L et al (2016) NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res 44:6614–6624
28. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

29. Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755
30. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG et al (2015) Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691–3693
31. Seemann T (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153
32. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780
33. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313
34. Sohail R, Jamil N. (2019) Isolation of biosurfactant producing bacteria from Potwar oil fields: effect of non-fossil fuel based carbon sources. Green Process Synth 9:77–86. https://doi.org/10.1515/gps-2020-0009
35 Alvarez VM, Jurelevicius D, Marques JM, de Souza PM, de Araújo LV, Barros TG et al (2015) Bacillus amyloliquefaciens TSBSO 3.8, a biosurfactant-producing strain with biotechnological potential for microbial enhanced oil recovery. Colloids Surf B Biointerfaces 136:14–21
36. Wu H, Wang D, Gao F (2021) Toward a high-quality pan-genome landscape of Bacillus subtilis by removal of confounding strains. Brief Bioinform 22:1951–1971. https://doi.org/10.1093/bib/bbaa013
37. Rasmussen S, Nielsen HB, Jarmer H (2009) The transcriptionally active regions in the genome of Bacillus subtilis. Mol Microbiol 73:1043–1057
38. Adékambi T, Drancourt M, Raoult D (2009) The rpoB gene as a tool for clinical microbiologists. Trends Microbiol 17:37–45
39. Ki J-S, Zhang W, Qian P-Y (2009) Discovery of marine Bacillus species by 16S rRNA and rpoB comparisons and their usefulness for species identification. J Microbiol Methods 77:48–57
40. Zhang S-J, Du X-P, Zhu J-M, Meng C-X, Zhou J, Zuo P (2020) The complete genome sequence of the algicidal bacterium Bacillus subtilis strain JA and the use of quorum sensing to evaluate its antialgal ability. Biotechnol Rep 25:e00421. https://doi.org/10.1016/j.btre.2020.e00421
41. Stice SP, Stumpf SD, Gitaitis RD, Kvitko BH, Dutta B (2018) Genetic diversity analysis reveals limited genomic diversity as well as accessory genes correlated with onion pathogenicity. Front Microbiol 9:184
42. Snipen L, Almøy T, Ussery DW (2009) Microbial comparative pan-genomics using binomial mixture models. BMC Genomics 10:385
43. Rouli L, Merhej V, Fournier P-E, Raoult D (2015) The bacterial pangenome as a new tool for analysing pathogenic bacteria. New Microbes New Infect 7:72–85
44. Coleman JJ, Rounsley SD, Rodriguez-Carres M, Kuo A, Wasmann CC, Grimwood J et al (2009) The genome of Nectria haematococca: contribution of supernumerary chromosomes to gene expansion. PLoS Genet 5:e1000618
45. Choi HJ, Shin D, Shin M, Yun B, Kang M, Yang H-J et al (2020) Comparative genomic and functional evaluations of Bacillus subtilis newly isolated from Korean traditional fermented foods. Foods 9:1805. https://doi.org/10.3390/foods9121805

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.