Functional & Integrative Genomics (2018) 18:533–543
https://doi.org/10.1007/s10142-018-0604-1

ORIGINAL ARTICLE

Genome annotation and comparative genomic analysis of *Bacillus subtilis* MJ01, a new bio-degradation strain isolated from oil-contaminated soil

Touraj Rahimi[1] · Ali Niazi[1] · Tahereh Deihimi[1] · Seyed Mohsen Taghavi[2] · Shahab Ayatollahi[3] · Esmaeil Ebrahimie[1,4,5,6]

Received: 7 December 2016 / Revised: 18 March 2018 / Accepted: 19 March 2018 / Published online: 5 May 2018

Abstract

One of the main challenges in elimination of oil contamination from polluted environments is improvement of biodegradation by highly efficient microorganisms. *Bacillus subtilis* MJ01 has been evaluated as a new resource for producing biosurfactant compounds. This bacterium, which produces surfactin, is able to enhance bio-accessibility to oil hydrocarbons in contaminated soils. The genome of *B. subtilis* MJ01 was sequenced and assembled by PacBio RS sequencing technology. One big contig with a length of 4,108,293 bp without any gap was assembled. Genome annotation and prediction of gene showed that MJ01 genome is very similar to *B. subtilis* spizizenii TU-B-10 (95% similarity). The comparison and analysis of orthologous genes carried out between *B. subtilis* MJ01, reference strain *B. subtilis* subsp. *subtilis str.* 168, and close relative spizizenii TU-B-10 by microscope platform and various bioinformatics tools. More than 88% of 4269 predicted coding sequences in MJ01 had at least one similar sequence in genome of reference strain and spizizenii TU-B-10. Despite this high similarity, some differences were detected among encoding sequences of non-ribosome protein and bacteriocins in MJ01 and spizizenii TU-B-10. MJ01 has unique nucleotide sequences and a novel predicted lasso-peptide bacteriocin; it also has not any similar nucleotide sequence in non-redundant nucleotide data base.

Keywords *Bacillus subtilis* · Whole genome · Biodegradation · Genome interpretation · Genomics comparison · Biosurfactant · Micro scope platform

Introduction

Leakage of crude oil and its derivatives to environment is one of the crucial contaminating factors of soil, air, and underground water (Bezza and Chirwa 2015; de Silva et al. 2014). Although some bacterial strains can degrade oil compounds, the low water solubility and hydrophobic characteristic of oil compounds cause low bio accessibility of these compounds for microbial digest (Liang et al. 2016).

Biosurfactant secretion is one of the main employed strategies in microorganisms for absorbing PAH aromatic hydro-carbons and hydrophobic compounds (Bezza and Chirwa 2015). *Bacillus subtilis* is an aerobic, rod-shaped, and GRAS (generally recognized as safe) bacterium (Sharma and Satyanarayana 2013). *B. subtilis* produces biosurfactant factors such as non-ribosome peptides (nrps) that is used for bioremediation of hydrocarbons (Bezza and Chirwa 2015) and improvement of enhanced oil recovery (Shibulal et al. 2014).

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10142-018-0604-1) contains supplementary material, which is available to authorized users.

✉ Ali Niazi
Niazi@shirazu.ac.ir

✉ Esmaeil Ebrahimie
EsmaeilEbrahimie@adelaide.edu.au

[1] Institute of Biotechnology, Shiraz University, Shiraz, Iran

[2] Department of Crop Protection, Shiraz University, Shiraz, Iran

[3] School of Chemical and Petroleum Engineering, Sharif University of Technology, Tehran, Iran

[4] School of Information Technology and Mathematical Sciences, Division of Information Technology, Engineering and the Environment, The University of South Australia, Adelaide, SA, Australia

[5] Adelaide Medicine School, The University of Adelaide, Adelaide, SA, Australia

[6] School of Biological Sciences, Faculty of Science and Engineering, Flinders University, Adelaide, SA, Australia

The produced peptide biosurfactant by *B. subtilis* has a range of activities from anti-microbial activities to eliminator agent in contaminated soils. Three main lipo-peptide compounds of surfactin, iturin, and fengycin families are produced by these bacterial strains (Ben Ayed et al. 2014). These molecules have various advantages compared to with chemical surfactants such as sustainability, lower toxicity, higher biodegradation capability, ecological adaptability, higher foam ability, higher selectivity, and specific activity. Furthermore, these strains can work on harsh conditions of high temperature, salinity, and pH (Ben Ayed et al. 2014; Bezza and Chirwa 2015; Jha et al. 2016).

B. subtilis has been also recognized as a model organism (Kamada et al. 2015); whole genome sequencing provided valuable information relating to biological functions, gene conservation, variation among specious and involved metabolic pathways for producing biosurfactants, and also oil bioremediation through sequence annotation (Sharma and Satyanarayana 2013).

The high performance of new generation of sequencing technology, its reasonable costs, and its higher efficiency compared to the first-generation sequencing have elaborated the insights into the bacterial whole genome sequencing (Kamada et al. 2014). This technology will be an important sequencing tool in the microbial genome studies (Land et al. 2015).

Pac Bio sequencing technology, the third generation creates long reads with relative length of 8500–30,000 bp, which facilitate the manipulation of these reads in complex regions such as repetitive elements. Therefore, genomes can be assembled with higher accuracy and validity by using long reads of Pac Bio. (Hutchison et al. 2016; Koren et al. 2013).

After accurate genome assembly, the homology of sequences and the information from the other reference genome and close relatives can improve the annotation (Ali et al. 2013). As example, the biochemical characteristics of the biosurfactants can be revealed from functional genomics analysis of available *Bacillus subtilis* genomes (Shaligram et al. 2016).

In this study, we isolated as a new strain of *B. subtilis* from oil-contaminated soil in south of Iran and its full genome was sequenced by PacBio technology. Then, assembling, annotation, and genome comparison analysis carried out based on coding sequences in MJ01 strain.

Material and methods

Growth conditions and preparation of genomic DNA

Bacillus subtilis MJ01 was grown aerobically in Luria Bertuni (LB) medium for 24 h under 35 °C and 260 rpm. Genomic DNA of this bacterium was extracted from LB medium using genomic DNA purification kit MG™ (Macrogen; Seoul, Korea).

Genome sequencing and assembling

To this end, 8 μl of purified genomic DNA was used using segmented g-TUBE and AMPure Bp magnetic willows. SMRTbell Template Prep Kit 1.0 was used to prepare library. Then, sequencing carried out based on PacBio RS system in MACROGEN Company (Seoul, Korea). Sequenced reads were filtered, mapped, and assembled by HGAP3 protocol and SMRT analysis v.2.3.0.140936 software (Rhoads and Au 2015).

Table 1 Statistical comparison of information and characteristics of genomes of *B. subtilis* MJ01, spizizenii TU-B-10, and 168 strains

Characteristic	No./amount MJ01	BSU168	BSUspizizenii TU-B-10
Contig No.	1	1	1
Nucleotide No.	4,108,293	4,215,606	4,207,222
GC%	43.93	43.51	43.82
Repeated zones-%	2.79	2.87	5.50
CDS mean length-bp	867.58	882.36	814.78
Mean length of gene gaps-bp	113.17	116.5	112.46
Concentration of coding protein-%	88.05	87.32	87.02
Total No. of genomic elements (miscRNA, tRNA, rRNA, fCDS, CDS)	4397	4468	4723
Total No of CDS	4269	4535	4617
Total No. of CDS without false genes	4218	4261	4601
Total No. of fCDS	26	51	40
Total misc-RNA	63	90	–
rRNA No. (including 5SrRNA, 23SrRNA, 16SrRNA)	30	30	30
tRNA No.	86	86	92

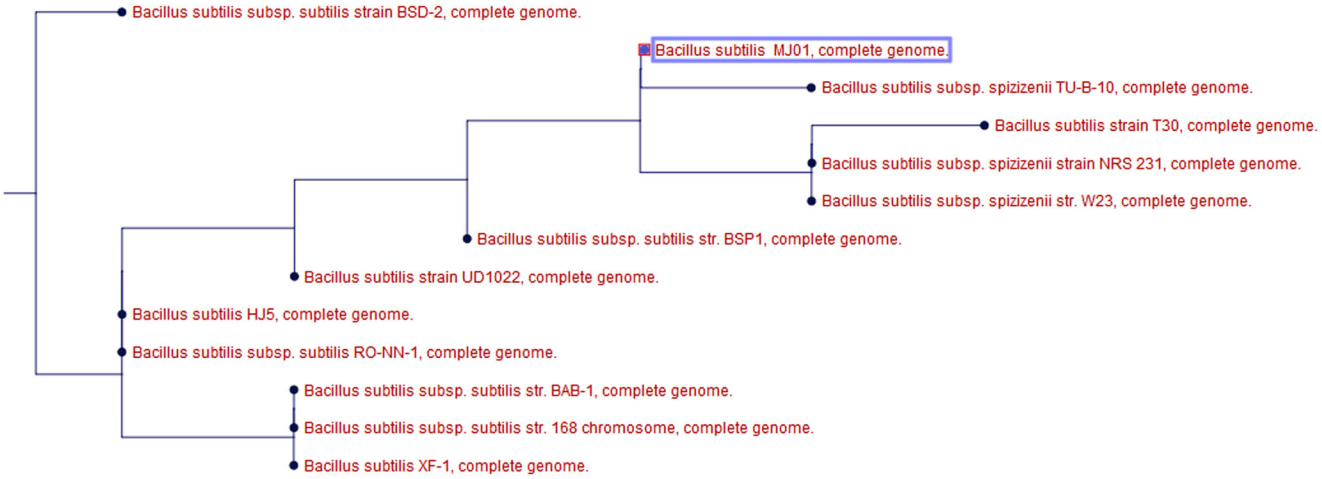

Fig. 1 Phylogenetic graph based on 16SrRNA sequence. The status of *B. subtilis* MJ01 is demonstrated versus the other relatives of *B. subtilis*

Submitting nucleotide sequences

The whole genome sequence of this bacterium was deposited in NCBI data base with accessibility Number CP-018173. The used version in this study is the first genome version.

Genome annotation, gene prediction, and coding zones

Genome interpretation, scanning, and gene prediction carried out by MicroScope platform (Vallenet et al. 2009; Vallenet et al. 2013). AntiSMASH v.3.05 (Weber et al. 2015) in MicroScope platform was used for identifying coding zones of secondary metabolites and non-ribosome peptides. In addition, BAGEL3 web-based database was used for predicting the coding sequences of Bacteriocin those assumed in *B. subtilis* MJ01 genome (Heel, Jong, Montalban-Lopez, Kok and Kuipers 2013). Resistance gene identifier (RGI) in CARD database was used for predicting coding sequences that are resistant to anti-biotic (McArthur and Wright 2015).

PHASTer was applied for recognition, interpretation, and indication of prophage sequences in MJ01 strains (Arndt et al. 2016). Moreover, genomic islands were detected by using online Web-based tool Island Viewer v.3 (Dhillon et al. 2015).

Comparative genomics analysis and protein classification

The comparative genomic sections of MicroScope platform such as pan and core genome and MAUVE alignment were used for comparing MJ01 genomes with reference genome strains of *B. subtilis* subsp. Subtilis str. 168 (NC-000964) and close relative *B. subtilis* subsp. spizizenii TU-B-10 (CP-002905). The classification of functional proteins carried out based on COG classification and total genome information by genomic tool of MicroScope platform. OrthoVenn platform was applied for comparing and interpreting orthologous gene cluster (Wang et al. 2015).

MJ01 genome was compared with whole genome of 45 strains of *B. subtilis* in NCBI database using GGDC v. 2.1 (Genome-Genome Distance calculator) online tools (Auch

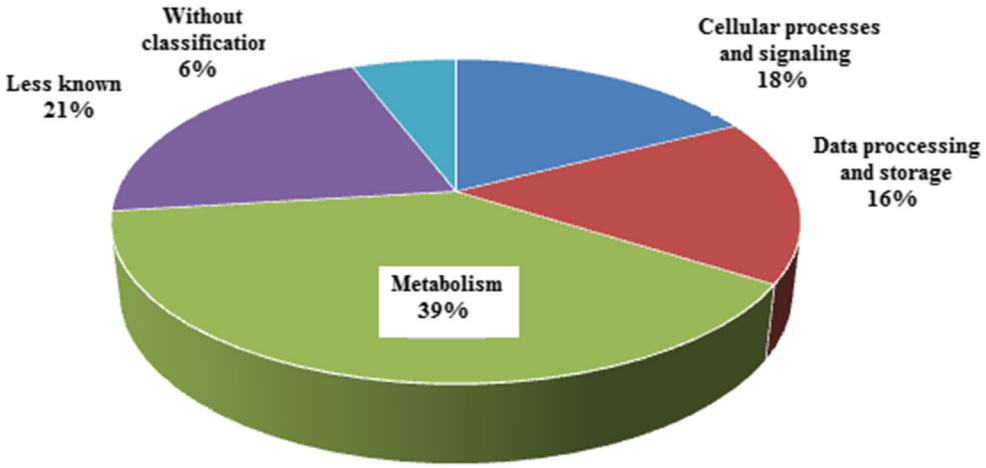

Fig. 2 Classification of functional proteins of MJ01 genome based on clusters of orthologus groups (COG)

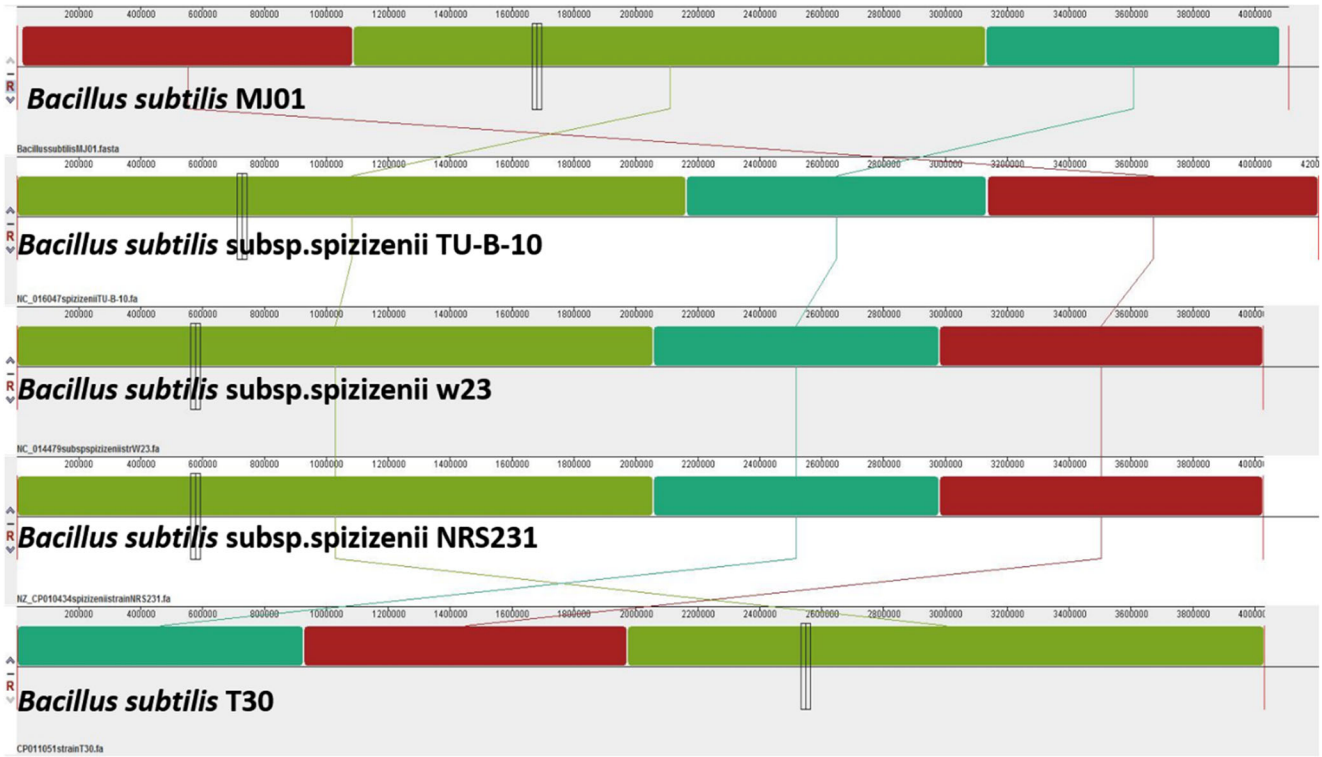

Fig. 3 Multi alignment of MJ01 genome compared to genomes of its close relatives *B. subtilis*, spizizenii TU-B-10, spizizenii W23, spizizenii NRS231 and T30. The same colored blocks mean conservation and homology between genomic zones

et al. 2010). Moreover, JSpecieWS was used to calculate ANI and four nucleotide correlation index (Richter et al. 2016).

Bacterial genome of MJ01 was entered to pubMLST online tool (http://pubmlst.org/) by using multi locus sequencing typing approach (MLST) for detecting taxonomic similarity in genetically loci of seven housekeeping genes (rpoD, tpiA, pycA, purH, glpF, pta, and ilvd) (Jolley and Maiden 2010).

Results

Genomic characteristics

MJ01 bacterial genome consists of a chromosome contig with sequence length of 4,108,293 bp with of 43.93% GC and 4269 coding sequences (Table 1). Table 1 shows total characteristics of MJ01 compared to spizizenii TU-B-10 and 168 strains. The length of MJ01 genome is 2% shorter than two other genomes of *B. subtilis*, while their GC percentages were relatively equal. Most of the strains of this group have 4 Mb length, and their GC percentages are between 43 and 44% (Fig. S1).

This genome includes 10 operons with 3 genes for rRNA and totally 30 rRNA genes were predicted in genome. The comparative consideration of rRNA operon numbers showed that the number of rRNA were similar for all three strains. In addition, 86 tRNA coding genes were detected for 20 standard amino acids on MJ01 chromosome. In terms of tRNA number, MJ01 genome was similar to 168 and both of them had 86 tRNA coding genes. In contrast, 90 sequences that code tRNA were observed in spizizenii TU-B-10.

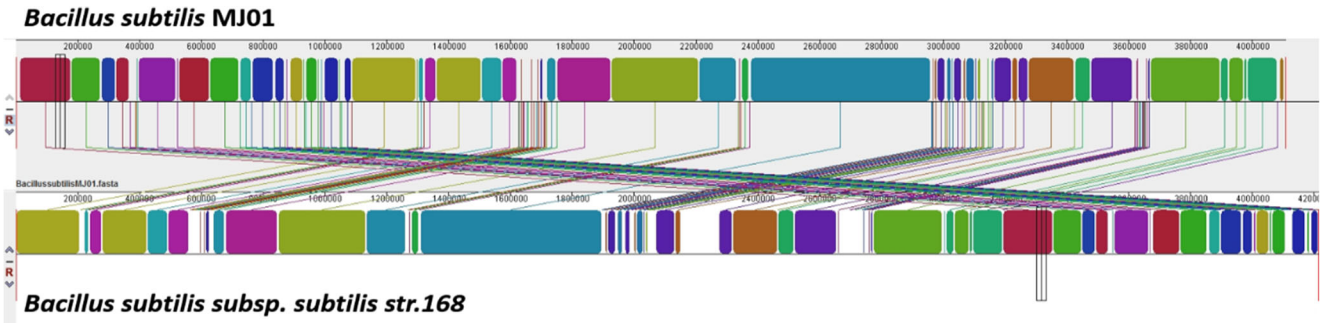

Fig. 4 Dual alignment of *B. subtilis* MJ01 and reference strain of *B. subtilis* subsp. *subtilis* str. 168 using Mauve software. The same colored blocks mean conservation and homology between genomic zones. Some parts of MJ01 and 168 genomes had not homology

Table 2 The results of searching profile of multi locus sequencing typing (MLST) in MJ01 genome of PubMLST data base

Isolate fields						MLST							
Id	Isolate	Name	Country	Specious	Subgroup	glpF*	ilvD	pta	purH*	pycA*	rpoD	tpiA	ST
not	MJ01	not	Iran	*Bacillus subtilis*	not	19	19	8	25	21	5	7	not
59	BGCS3A17	N10	not	*Bacillus subtilis*	spizizenii	19	19	8	25	21	5	21	15

Based on search of the best sequence, the number of selected locus for MJ01 genome has been selected as assorted alignment. Not means not defined or no data available

The analysis of 16SrRNA sequence showed that MJ01 bacterium was 100% similar to 16SrRNA sequence of *B. subtilis* strain AER314-2; on the other hand, *B. subtilis* subsp. Spizizenii TU-B-10, *Bacillus* sp. JS, and *B. subtilis* strain BS3902 were similar (Fig. 1).

COG (clusters of orthologus groups) analysis was classified 3972 proteins out of 4218 protein coding sequences, which were predicted in MJ01 genome (Fig. 2). COG class divided coding proteins of MJ01 to four main groups of signaling and cell processes, information processing and storage, metabolism and fewer known and 21 classes (Table S1). The most coding sequences (about 39%) were grouped in metabolism class. MJ01 bacterium is the producer genome of secondary metabolites similar to the other strains of *B. subtilis*. Therefore, the presence of sequences (107 coding sequences) that produce secondary metabolites in genome of this bacterium is significant (about 2.57%).

Comparative genomics analysis of MJ01 strain with other *B. subtilis* strains

Whole-genome alignment by using MAUVE

Whole-genome alignment carried out for four genomes with the most similarity level, that carried out by BLAST search of MAUVE genome alignment software V. 20150226 and included spizizenii TU-B-10, W23, NRS 231, and T30 and MJ01 strains (Fig. 3). Three conserve blocks were detectable in full genome alignment.

Furthermore, MJ01 genome was aligned with the genome of 168 strain as reference genome of subtilis subgroup. The result of alignment of these two genomes showed that although there are conservative blocks between two genomes, non-homolog zones exist between two genomes (Fig. 4).

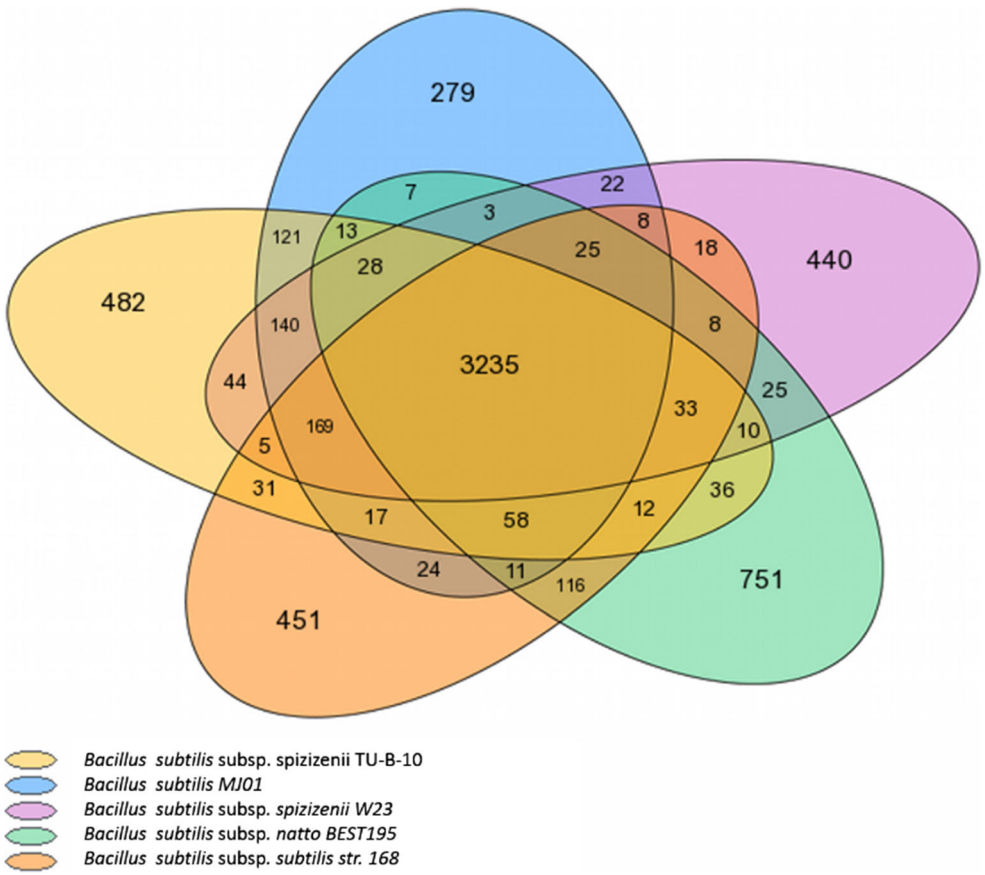

Fig. 5 Pan-genome analysis of *B. subtilis* close relatives with MJ01 genome in MicroScope platform. Distribution of gene families in the core and strain-specific genome was 70/30

Table 3 The number of genes involved in pan-genome analysis of five relative bacterial strains. The results show that *B. subtilis* MJ01 genome has fewer strain-specific gene

Strain name	Number of CDS	Pan CDS	Core CDS	Variable CDS	Strain-specific CDS	Percentage of core CDS	Percentage of variable CDS	Percentage of strain–specific CDS	Out of analysis CDS
Bacillus subtilis MJ01	4218	4187	3254	933	279	71.71	22.28	2.66	0
Bacillus subtilis subsp. subtilis str. 168	4261	4261	3265	996	451	76.63	23.37	10.58	0
Bacillus subtilis subsp. spizizenii W23	4284	4250	3253	997	440	76.54	23.45	10.35	0
Bacillus subtilis subsp. natto BEST195	4533	4452	3252	1200	785	73.04	26.95	17.63	0
Bacillus subtilis subsp. spizizenii TU-B-10	4601	4554	3276	1278	537	71.94	28.06	11.79	0

Genome alignment shows that organization of MJ01 genome and its conservative zones are similar to strains in spizizenii group.

Hybridization of DNA-DNA by using GGDC

Genome-to-genome distance calculate based amount of digital DNA-DNA hybridization (DDH) showed that MJ01 genome had the most DDH with the genomes of bacteria that are subgroup of spizizenii TU-B-10, W23, and NRS231 strains at the level of formula 1, 94.7, 93.6, and 93.6%, respectively (Table S2).

Average nucleotide identity

Analysis of four nucleotide correlations among all whole genomes and contings of MJ01 in JSpeciesws database (Richter et al. 2016) shows that the genome of MJ01 bacterium had correlation with 15 genomes with more than 0.999 in z-score domain (Table S3). The genome of *B. subtilis subps.* Spizizenii TU-B-10 with z-score = 0.9999 had the most correlation. Consequently, the contig of *Jeotagalibacillus marinus* DSM 1297 (separated from unknown sediment source) is located at the second rank and after that the contig of *B. subtilis* JRS7 (separated from dessert soil) with the most z-score (0.99959 and 0.9995, respectively). Among completed genomes, the most z-score was calculated for *B. subtilis* subsp. spizizenii W23 and NRS 231 (z-score = 0.99946 for both of them). The results of two methods of nucleotide similarity calculation (based on ANIb or BLAST and ANIm or Mummer) show that MJ01 had the most ANIb with the genome of bacteria that are subgroup of spizizenii such as TU-B-10, NRS 231, and W23, 99.13, 96.52, and 96.52%, respectively. In addition, ANIm amounts were 99.25, 96.74, and 96.73 for these strains. ANI analysis confirmed that the MJ01 genome can be located as a subgroup of spizizenii bacteria.

Multi locus sequence typing

Search of pubMLST data base for locus sequence of seven housekeeping genes *rpoD*, *ilvD*, *pta*, *purH*, *pycA*, *glpF*, and *tpiA* for *B. subtilis* in MJ01 genome shows that three genes of *pta*, *ilvD*, and *rpoD* had similar locus with same directions. Gene locus of *tpiA*, *pycA*, *purH*, and *glpF* did not follow same direction in the database. After considering locus

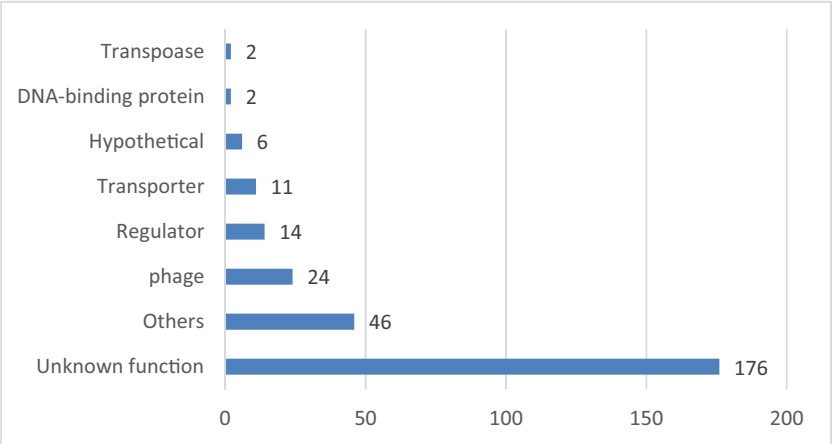

Fig. 6 Classification of strain-specific genes for *B. subtilis* MJ01 strain. The result showed that the source of most of these genes is unknown and some of them are the results of horizontal transference of gene through prophages or transposon elements

Table 4 The result of analyzing *B. subtilis* MJ01 genome by using AntiSMASH online tool that predicts coding sequences of secondary metabolites

Start	Stop	Length	Cluster type	Compound of peptide monomers	The name of secondary metabolite
103,816	113,562	9747	NRPS	(dhb) + (gly-thr)	Bacillibactin
381,631	383,903	2273	lassopeptide	–	putative Asparagine synthase (Glutamine-hydrolyzing)
448,317	449,063	747	Other	–	cyclodipeptide synthase
702,057	703,403	1347	Sactipeptide	–	Subtiliosin_A
737,237	738,655	1419	Other	–	Bacilysin
1,458,804	1,484,191	25,391	NRPS	(glu-leu-leu) + (val-asp-leu) + (leu)	Surfactin
2,234,503	2,235,309	807	terpene	–	farnesyl diphosphate phosphatase
2,853,896	2,923,818	69,923	otherks-nrps-transatpks	(mal) + (pk) + (mal) + (nrp-gly) + (nrp)	Bacillaene
3,020,638	3,063,549	42,912	nrps-transatpks	(mal) + (pk-asn) + (tyr-asn-gln-pro) + (ser-asn) + (nrp)	Mycosubtilin
3,193,315	3,195,213	1899	terpene	–	squalene-hopene cyclase
3,273,643	3,274,740	1098	t3pks	–	promiscuous alkylpyrone synthase *BpsA*
103,816	113,562	9747	NRPS	(dhb) + (gly-thr)	Bacillibactin

sequencing typing of housekeeping genes on MJ01 genome by online tool in pubMLST web site, the result confirmed that the profile of MJ01 is exclusive to this bacterium and it has not been recorded in pubMLST web site so far. Therefore, the most similar profile belongs to *B. subtilis* bacterium subgroup of spizizenii and BGSC3A17 strain isolated No. 10 (Table 2).

Pan and core genome

Pan and core analysis of four genomes of natto BEST 195, W23, TU-B-10, and 168 strains along with MJ01 in MicroScope platform (Fig. 5) demonstrates that interpreted genes in 5 genomes have formed a gene pool with 21,704 genes. These genes have

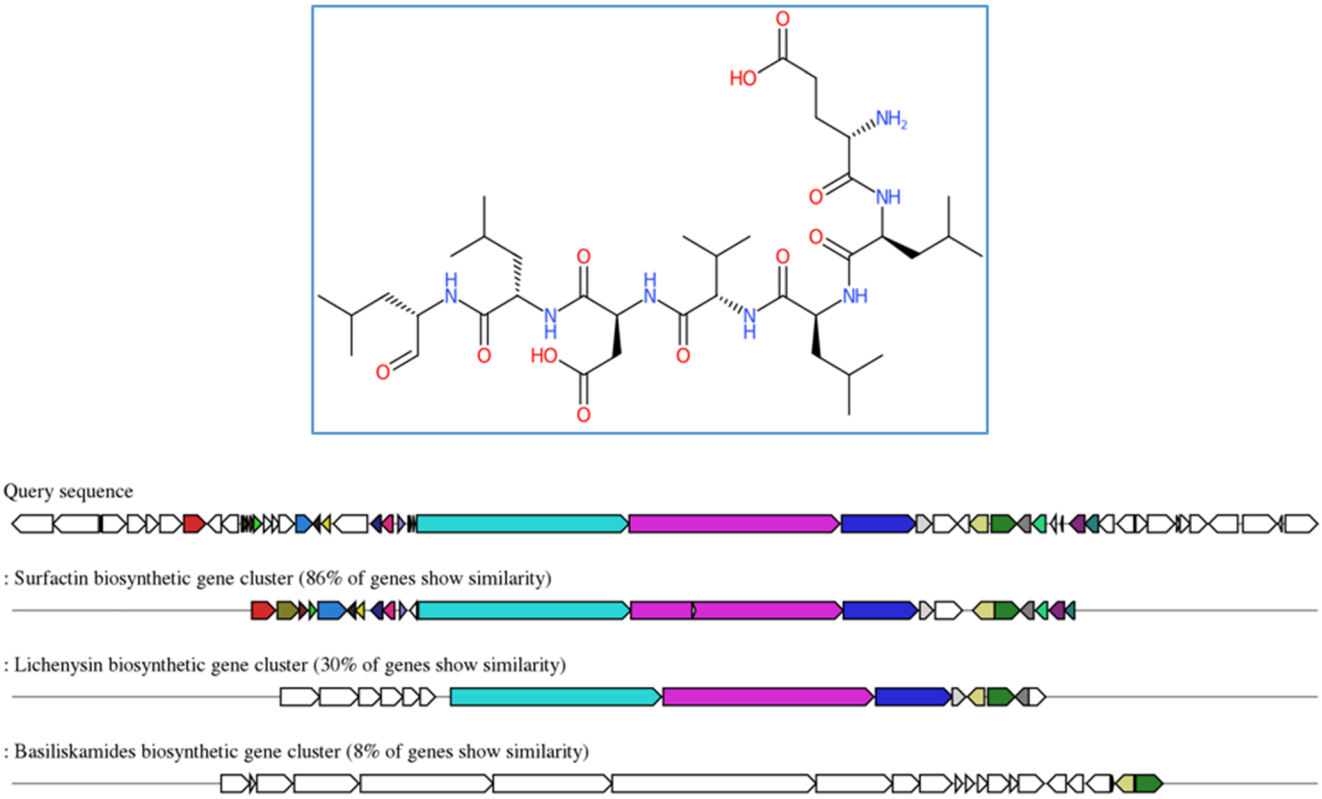

Fig. 7 The cluster of gene that biosynthesizes surfactin lipo peptide and its predicted structure in MJ01 and its comparison with gene cluster of other bacterial specious existing in the data base

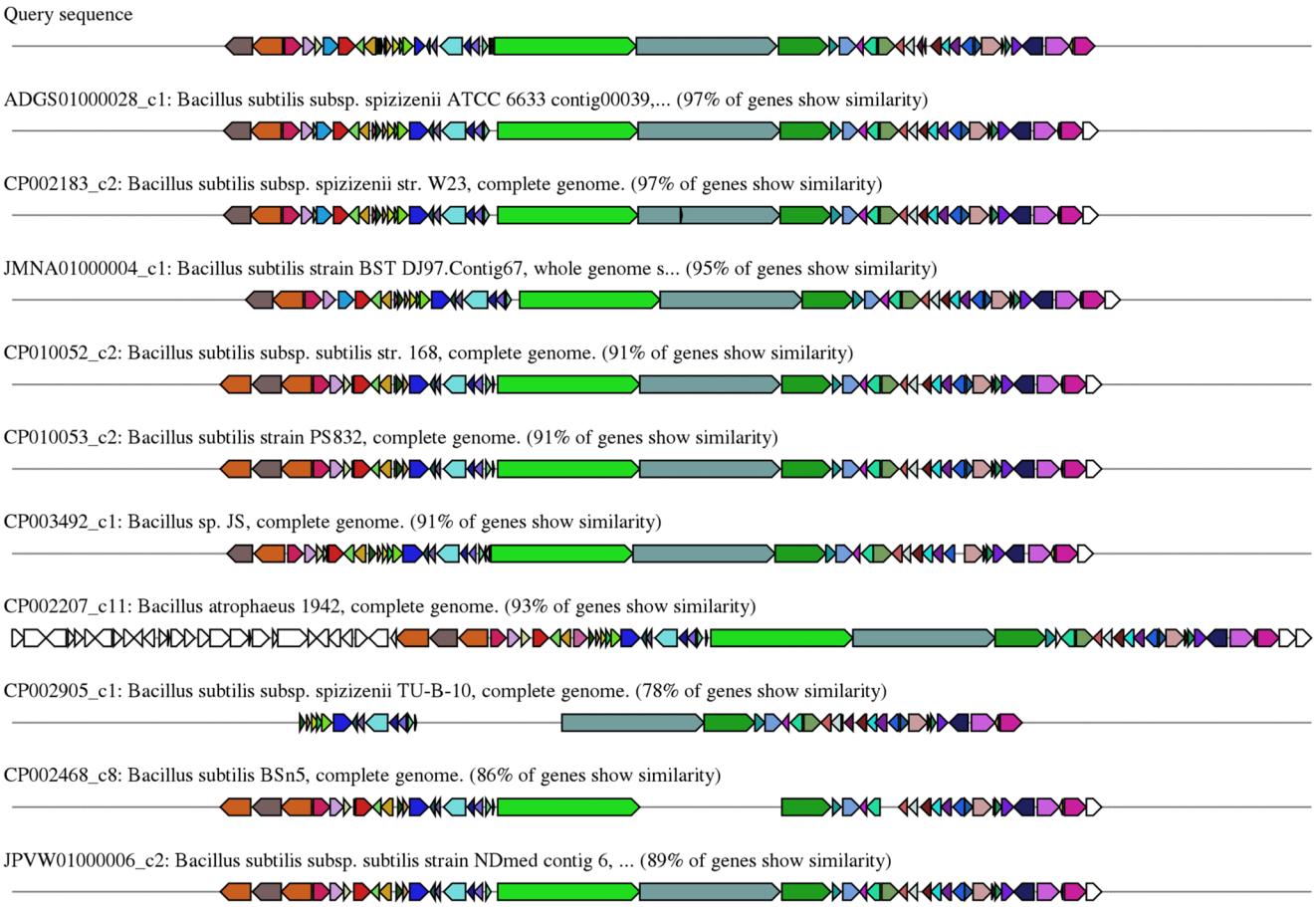

Fig. 8 The comparison of gene clusters for surfactin lipo peptide biosynthesis between *B. subtilis* MJ01 and other bacterial strains. The conservative *synteny* zones are observable comparing with area of interest and aimed sequences. Different genes are determined with different colors and homolog genes have same colors

been classified into 6622 gene family based on 80% similarity of amino acid and 80% alignment coverage (Table 3).

MJ01 strain-specific genes

Genome pan/core analysis showed that 279 coding sequences that are strain-specific have been detected in MJ01 genome. They include 2.66% of coding sequences in MJ01 genome. Fewer strain-specific genes in MJ01 compared to the other strains. The fewer amounts of strain-specific genes are not out of expectation according to shorter length of the genome and fewer predicted coding sequences. These strain-specific genes included phage, regulator, and unknown proteins (Fig. 6).

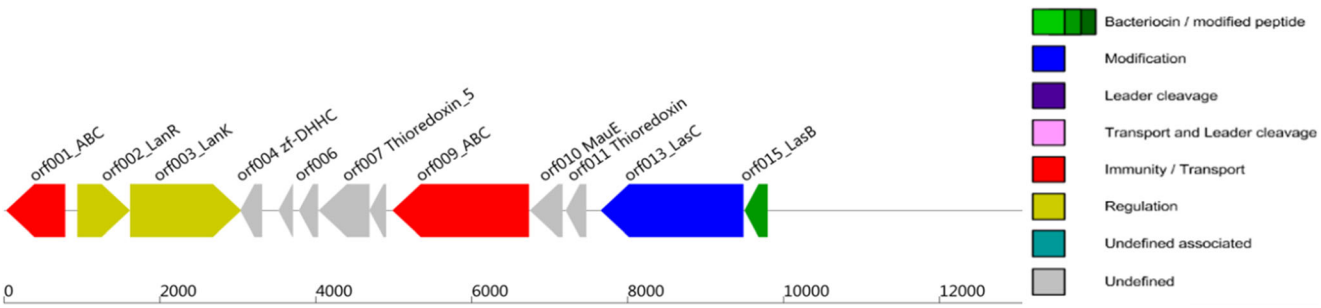

Fig. 9 The predicted areas that prone to produce bacteriocin in MJ01 genome by BAGEL3 on line tool. Coding sequences of bacteriocin (green), transporter (red), adjusted areas (yellow), and rectifier section (blue) have been shown

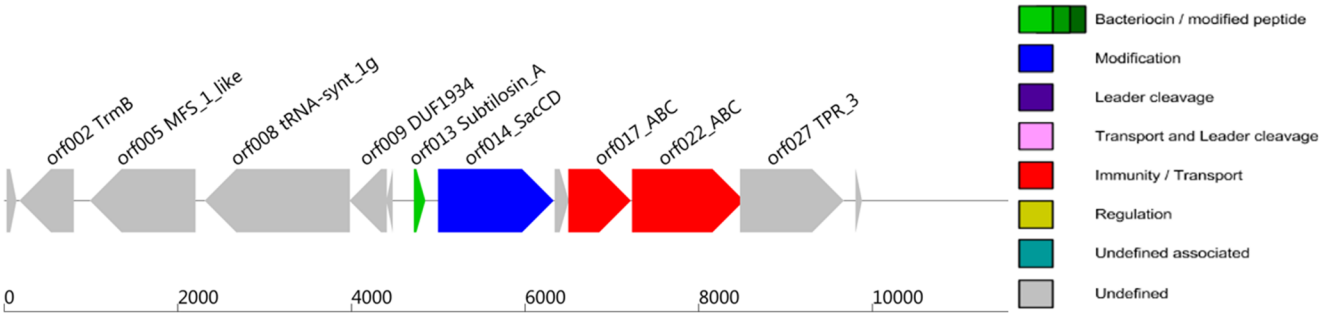

Fig. 10 The predicted areas that prone to produce subtilosin A bacteriocin n MJ01 genome by BAGEL3 on line tool. Coding sequences of bacteriocin (green), transporter (red), and rectifier section (blue) have been shown

Non-ribosome proteins and anti-biotic coding sequences

Detection and prediction of secondary metabolites and non-ribosome lipo peptide

The results of MJ01 genome analysis by AntiSMASH on line tool V3.0.5, 11 gene zones for production of secondary metabolites, and non-ribosome lipo proteins are presented in Table 4. Four gene zones that were responsible for coding non-ribosome lipo proteins (NRPS) in MJ01 genome were detected including Bacillibactin, Surfactin, Bacilliacne, and Mycosubtilin. Moreover, other vulnerable sections that coding NRPS were detected such as Basilysin in position of 737,237 to 738,655 and a section that codes polyketide in positions of 3,273,643 to 3,274,740. Two involved gene zones in synthesis of lassopeptide and sactipeptide bacteriocins classes were predicted in the genome of MJ01 bacterium. In addition, there are two zones that produce terpenes in this genome.

Surfactin coding operon locates on the positions of 1,458,804 to 1,484,194 of MJ01 genome. This operon is similar to gene cluster of data base in terms of gene sequence (86%) (Figs. 7 and 8). Studying homolog gene cluster of this operon on relative bacteria showed that this operon has the most similarity (97%) with its homolog in *B. subtilis* subsp. spizizenii W23 bacteria. The homolog similarities of this operon with 168 and spizizenii TU-B-10 (the closest relative) are 91 and 78% in order (Fig. 7).

According to the obtained data from AntiSMASH and annotation of MJ01 genome, the position of *SrfA* operon and *sfp* gene were detected on the genome. This operon has 4 genes of *srfA*-ABCD in positions of 1,458,804–1,484,950. *Sfp* gene locates on 5017 bp at lower part of surfactin operon that biosynthesizes surfactin.

Mycosubtilin producer operon has been predicted on the positions of 3,020,638–3,063,549 with the length of 42,912 bp on MJ01 genome. The seeking on the sequence of this gene cluster in the data base shows that it has 100% similarity with mycosubtilin gene cluster. This gene cluster is 100% similar with its homolog in close relative bacterium (spizizenii TU-B-10), and it is also 94% similar with reference genome of 168. By using achieved data from AntiSMASH analysis and interpretation of MJ01 genome, the exact position of this operon was detected on the genome. This operon has four genes of *myc*A, *myc*B, *myc*C, and *fen*F on the positions of 3,027,547–3,064,769 and a negative strand.

Entrance of MJ01 genome to BAGLE3 online tool predicts two areas of interest for accepted bacteriocins. The first area locates in lasso-peptide class and position of 373,960–383,960 (Fig. 9). The BLAST search of this 10,000 nucleotide area in NCBI data base shows that there is not any nucleotide sequence in the data base that is significantly similar with it. The result of search is only a sequence including 28 nucleotides of *Galdieria sulphuraria* H+-translocating *PPase* (*vocuolar*) (Gasu-15,749), mRNA. *Galderia* is a unicellular red alga. It seems that the BLAST result of nucleotides in this predicted area indicates that this sequence specifies to the genome of MJ01 and a new bacteriocin.

The second area locates in sactipeptides class and positions of 697,075–707,075 on MJ01 genome (Fig. 10). The nucleotide BLAST search of this area in NCBI data base shows that this area has nucleotide similarity in the genomes of *B. subtilis*

Table 5 Prophage prone areas in *B. subtilis* MJ01 genome; two incomplete prophage areas and a relatively complete prophage area in *B. subtilis* MJ01 genome have been predicted by Phaster on line tool

Status of area on genome	Numbers of proteins	The completeness level	Area size Kb	GC %	Related phage	Score
363–16,678	29	Incomplete	16.3	43.53	PHAGE_Brevib_Davies_NC_022980	20
2,382,097–2,416,062	46	Treatable	33.9	44.87	PHAGE_Brevib_Jimmer1_NC_029104	90
4,083,170–4,107,066	34	Incomplete	23.8	42.11	PHAGE_Brevib_Jimmer1_NC_029104	40

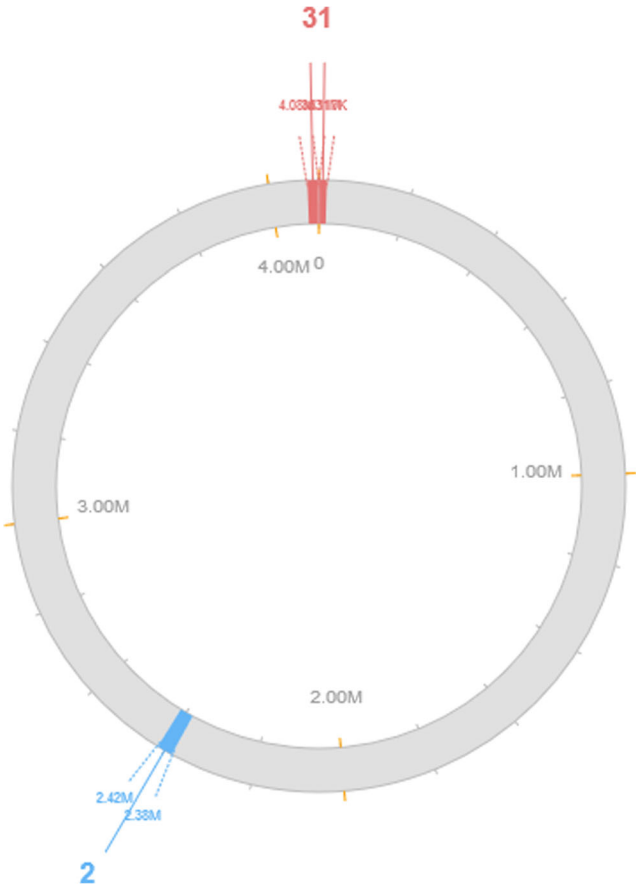

Fig. 11 The view of detected prophage areas in MJ01 genome by phaster on line tool. There are two incomplete prophage areas (red) and a relatively complete prophage area (blue) in the structure of MJ01 genome

subsp. Spizizenii TU-B-10, *B. subtilis* T30, *B. subtilis* subsp. spizizenii NRS231 and W23; the similarities are 99, 97, 97, and 97% with 100% coverage.

Resistance to anti-biotic

The search on MJ01 genome was performed to find involve genes that are resistant to anti-biotic. This search carried out by RGI (resistance gene identifier) through CARD on microscope platform that has led to detection of 22 coding sequence in the genome and 19 CDS that had homology with other existed genes in the data base and 3 CDS were identified as strain, which had single nucleotide polymorphism (SNP) in CARD database. Among these coding sequences, there are membrane pumps for resisting against peptide anti-biotics such as *lmrB* and *mprF*; in addition, *ykkD* and *bmr* genes were effective on resistance to Chloramphenicol anti-biotic and *penR* gene was effective on hydrolysis of beta Lactam chain of penicillin (Table S4).

Prophage sequences existed in *B. subtilis* MJ01 genome

Based on the searches that were carried out by PHASTer on line tool, three prophage areas are observable in MJ01 genome. One area includes a prophage section with the score of more than 70 and less than 90 and two areas including incomplete prophage genes (Table 5, Fig. 11).

Discussion

The long read sequences of *Bacillus subtilis* MJ01 genome that were prepared by PacBio RS SMRT provide high quality and accuracy of genome assembly and show success in a creation of a scaffolding circular chromosome. This high quality is the result of applying proper platform with preferable and long reading size more than 10,000 nucleotides and the absence of vague bases. Therefore, no gap has been created for contig assembly and genome scaffolding. Therefore, there is no need to use common methods such as PCR for filling the gaps, contrasting to the common studies that use short reads of other platforms (Koren et al. 2013).

Analysis in genome scale provides information about genome organization and its similarity with the other bacterial relative strains of *B. subtilis* group. Based on these analyses, it was determined that total organization of MJ01 genome is very similar to subgroups of spizizenii. Obtained information and results from studying whole alignment of genome that carried out by Mauve tool, DDH method, and ANI calculation have confirmed this point. Multi locus sequencing typing (MLST) results at the level of housekeeping genes and profile analysis showed that although there are close similarities between *B. subtilis* MJ01 genes and their homologs in *B. subtilis* spizizenii TU-B-10, the structure of MJ01 has been subjected to events such as point mutations and inversion. This issue not only caused differentiation of MJ01 bacterium from its close relatives such as spizizenii-TU-B-10 but also it confirmed that MJ01 is a unique strain and probably a subgroup of spizizenii.

Distribution of gene families in analysis with relative ratio of 70 to 30 among core genome and strain-specific genome was not in line with the findings of Yu et al. (2015). They used 13 genomes of *B. subtilis* and analyzed pan-genome to find that distribution of gene families follows a balanced ratio of 50/50 among core genome and strain-specific genome; they concluded that existed genes in core genome of *B. subtilis* were under higher purification selective pressure than strain-specific genes (Yu et al. 2015).

Ortholog protein clustering and gene classification based on pan-core have detected unique functional coding sequences in MJ01 genome. The searches of these unique sequences in non-repeated nucleotide sequence database showed that under studied sequence has not any similar sequence in the database. Therefore, the studies on new sequences can draw attentions toward interpretation of newfound genes of *B. subtilis* bacteria genetics in future, and also can provide better insight through genomic information about similar metabolic activities in relative bacteria that have similar genetic machine and coding sequences (Harvey et al. 2015).

In conclusion, using genomic comparative attitude has indicated that MJ01 genome has similar genomic structure with spizizenii-TU-B-10. However, MJ01 genome has distinctive differences with spizizenii-TU-B-10 strain such as new bacteriocin coding sequence.

Acknowledgements We thank to the LABGeM and the National Infrastructure « France Genomique », for their useful MicroScope platform tools and providing genome annotation and comparative analysis for MJ01 genome. We would also thanks to Mr. Moien Jahanbani Veshareh for providing MJ01 strain bacteria. This study was supported by Department of Biotechnology, Agriculture Faculty of Shiraz University.

References

Ali A, Soares SC, Barbosa E, Santos AR, Barh D, Bakhtiar SM et al (2013) Microbial comparative genomics: an overview of tools and insights into the genus Corynebacterium. J Bacteriol Parasitol 4(2): 1–16. https://doi.org/10.4172/2155-9597.1000167

Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS (2016) PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res 1–6. https://doi.org/10.1093/nar/gkw387

Auch AF, von Jan M, Klenk H-P, Göker M (2010) Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand Genomic Sci 2(1):117–134. https://doi.org/10.4056/sigs.531120

Ben Ayed H, Hmidet N, Béchet M, Chollet M, Chataigné G, Leclère V, Jacques P, Nasri M (2014) Identification and biochemical characteristics of lipopeptides from Bacillus mojavensis A21. Process Biochem 49(10):1699–1707. https://doi.org/10.1016/j.procbio.2014.07.001

Bezza FA, Chirwa EMN (2015) Production and applications of lipopeptide biosurfactant for bioremediation and oil recovery by Bacillus subtilis CN2. Biochem Eng J 101:168–178. https://doi.org/10.1016/j.bej.2015.05.007

de Silva R, CFS, Almeida DG, Rufino RD, Luna JM, Santos VA, Sarubbo LA (2014) Applications of biosurfactants in the petroleum industry and the remediation of oil spills. Int J Mol Sci Multidiscip Digit Publ Inst (MDPI) 15:12523–12542. https://doi.org/10.3390/ijms150712523

Dhillon BK, Laird MR, Shay JA, Winsor GL, Lo R, Nizam F, Pereira SK, Waglechner N, McArthur AG, Langille MGI, Brinkman FSL (2015) IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. Nucleic Acids Res 43(W1):W104–W108. https://doi.org/10.1093/nar/gkv401

Harvey AL, Edrada-Ebel R, Quinn RJ (2015) The re-emergence of natural products for drug discovery in the genomics era. Nat Rev Drug Discov 14(2):111–129. https://doi.org/10.1038/nrd4510

Hutchison CA, Chuang R-YR-Y, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH et al (2016) Design and synthesis of a minimal bacterial genome. Science 351(6280):aad6253–aad6253. https://doi.org/10.1126/science.aad6253

Jha SS, Joshi SJ, Geetha SJ (2016) Lipopeptide production by Bacillus subtilis R1 and its possible applications. Braz J Microbiol 47(4): 955–964. https://doi.org/10.1016/j.bjm.2016.07.006

Jolley KA, Maiden MCJ (2010) BIGSdb: scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11(1):595. https://doi.org/10.1186/1471-2105-11-595

Kamada M, Hase S, Sato K, Toyoda A, Fujiyama A, Sakakibara Y (2014) Whole genome complete resequencing of Bacillus subtilis natto by combining long reads with high-quality short reads. PLoS One 9(10):e109999. https://doi.org/10.1371/journal.pone.0109999

Kamada M, Hase S, Fujii K, Miyake M, Sato K, Kimura K, Sakakibara Y (2015) Whole-genome sequencing and comparative genome analysis of bacillus subtilis strains isolated from non-salted fermented soybean foods. PLoS One 10(10):e0141369. https://doi.org/10.1371/journal.pone.0141369

Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, Mcvey SD, Radune D, Bergman NH, Phillippy AM (2013) Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol 14(9):R101. https://doi.org/10.1186/gb-2013-14-9-r101

Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW (2015) Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics 15(2):141–161. https://doi.org/10.1007/s10142-015-0433-4

Liang Y, Zhao H, Deng Y, Zhou J, Li G, Sun B (2016) Long-term oil contamination alters the molecular ecological networks of soil microbial functional genes. Front Microbiol 7:60. https://doi.org/10.3389/fmicb.2016.00060

McArthur AG, Wright GD (2015, October) Bioinformatics of antimicrobial resistance in the age of molecular epidemiology. Curr Opin Microbiol 27:45–50. https://doi.org/10.1016/j.mib.2015.07.004

Rhoads, A., & Au, K. F. (2015). PacBio sequencing and its applications. Genomics, Proteomics and Bioinformatics, pp. 278–289. 10.1016/j.gpb.2015.08.002, PacBio Sequencing and Its Applications

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(6):929–931. https://doi.org/10.1093/bioinformatics/btv681

Shaligram S, Kumbhare SV, Dhotre DP, Muddeshwar MG, Kapley A, Joseph N, Purohit HP, Shouche YS, Pawar SP (2016) Genomic and functional features of the biosurfactant producing Bacillus sp. AM13. Functional and Integrative Genomics 1–10. https://doi.org/10.1007/s10142-016-0506-z

Sharma A, Satyanarayana T (2013) Comparative genomics of Bacillus species and its relevance in industrial microbiology. Genomics Insights Libertas Academica 6:GEI.S12732. https://doi.org/10.4137/GEI.S12732

Shibulal B, Al-Bahry SN, Al-Wahaibi YM, Elshafie AE, Al-Bemani AS, Joshi SJ (2014) Microbial enhanced heavy oil recovery by the aid of inhabitant spore-forming Bacteria: an insight review. Sci World J 2014:1–12. https://doi.org/10.1155/2014/309159

Vallenet D, Engelen S, Mornico D, Cruveiller S, Fleury L, Lajus A et al (2009) MicroScope: a platform for microbial genome annotation and comparative genomics. Database 2009:bap021. https://doi.org/10.1093/database/bap021

van Heel AJ, de Jong A, Montalbán-López M, Kok J, Kuipers OP (2013) BAGEL3: automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. Nucleic Acids Res 41(Web Server issue):W448–W453. https://doi.org/10.1093/nar/gkt391

Vallenet D, Belda E, Calteau A, Cruveiller S, Engelen S, Lajus A, le Fèvre F, Longin C, Mornico D, Roche D, Rouy Z, Salvignol G, Scarpelli C, Thil Smith AA, Weiman M, Médigue C (2013) MicroScope–an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. Nucleic Acids Res 41(Database issue):D636–D647. https://doi.org/10.1093/nar/gks1194

Wang Y, Coleman-Derr D, Chen G, Gu YQ (2015) OrthoVenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Res 43(W1):W78–W84. https://doi.org/10.1093/nar/gkv487

Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY, Fischbach MA, Müller R, Wohlleben W, Breitling R, Takano E, Medema MH (2015) antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res 43(W1):W237–W243. https://doi.org/10.1093/nar/gkv437

Yu G, Wang XC, Tian WH, Shi JC, Wang B, Ye Q, Dong SG, Zeng M, Wang JZ (2015) Genomic diversity and evolution of Bacillus subtilis. Biomed Environ Sci : BES 28(8):620–625. https://doi.org/10.3967/bes2015.087