Whole Genome Sequence of biosurfactant producing *Bacillus tequilensis*

Anuraj Nayarisseri¹,²,³,*; Sanjeev Kumar Singh¹*

*Corresponding authors Email: anuraj@eminentbio.com; skysanjeev@gmail.com

1. Computer Aided Drug Designing and Molecular Modeling Lab, Department of Bioinformatics, Alagappa University, Karaikudi-630 003, Tamil Nadu, India.
2. In silico Research Laboratory, Eminent Biosciences, Indore – 452 010, Madhya Pradesh, India.
3. Bioinformatics Research Laboratory, LeGene Biosciences Pvt Ltd., Mahalakshmi Nagar, Indore - 452010, Madhya Pradesh, India.

*Corresponding authors:

Anuraj Nayarisseri
Email: anuraj@eminentbio.com
Computer Aided Drug Designing and Molecular Modeling Lab, Department of Bioinformatics, Alagappa University, Karaikudi-630 003, Tamil Nadu, India.
Madhya Pradesh, India.
Tel: +91 9752295342

Prof. Dr. Sanjeev Kumar Singh
Email: skysanjeev@gmail.com
Affiliation: Computer Aided Drug Designing and Molecular Modeling Lab, Department of Bioinformatics, Alagappa University, Karaikudi-630 003, Tamil Nadu, India.

ORCID:
Anuraj Nayarisseri: [http://orcid.org/0000-0003-2567-9630](http://orcid.org/0000-0003-2567-9630)
Sanjeev Kumar Singh: [http://orcid.org/0000-0003-4153-6437](http://orcid.org/0000-0003-4153-6437)

Abstract

**Background:** Bioremediation is crucial for recuperate polluted water and soil. By expanding the surface area of substrates, biosurfactants play a vital role in bioremediation. Bioremediation is crucial for recuperate polluted water and soil. By expanding the surface area of substrates, biosurfactants play a vital role in bioremediation. Biosurfactant producing microbes release certain biosurfactant compounds, which are promoted for oil spill remediation. In the present investigation, a biosurfactant producing bacterium *Bacillus tequilensis* was isolated from Chilika lake, Odisha, India (latitude and longitude: 19.8450 N 85.4788 E). Whole Genome Sequencing (WGS) of *Bacillus tequilensis* was carried out using Illumina NextSeq 500.
Results: The whole genome sequence is 4.47 MB consisting of 4,478,749 base pairs forming a circular chromosome with 528 scaffolds, 4492 protein-encoding genes (ORFs), 81 tRNA genes, and 114 ribosomal RNA transcription units. The total number of raw reads was 4209415, and processed reads were 4058238 with predicted genes of 4492. The whole-genome obtained from the present investigation was used for genome annotation, variant calling, variant annotation and comparative genome analysis with other existing Bacillus species. In this study, a pathway was constructed which describe the biosurfactant metabolism of Bacillus tequilensis. The study identified genes such as SrfAD, SrfAC, SrfAA, SrfAB which are involved in biosurfactant synthesis.

Conclusion: The sequence of the genes SrfAD, SrfAC, SrfAA, SrfAB was deposited in Genbank database with accession MUG02427.1, MUG02428.1, MUG02429.1, MUG03515.1 respectively. The whole-genome sequence was submitted to Genbank with an accession RMVO0000000 and the raw reads can be obtained from SRA, NCBI repository using accession: SRX5023292.

Keywords: Biosurfactant producing bacteria, Bacillus tequilensis, Genome sequence, Genome annotation, Next Generation Sequencing, De-novo assembly, Biosurfactant producing genes.

1. INTRODUCTION

Heavy metal contamination has now become serious ecological threat raising environmental concerns. Metals especially cadmium and zinc are has posed serious threat as their degradation to innocuous products is hard and takes millions of years [1-3]. Bioremediation systems have however been long proposed to neutralize metal contamination, however, have low bioavailability leading to incomplete bioremediation process. Further, such bioremediation process like phytoremediation with synthetic chelators have been shown be expensive and environmentally hazardous [4-5]. Various surface-active compounds (SACs) commonly the biosurfactants produced by microorganisms have emerged as safe alternative to chemical remediation that are known to be safe, and are now exploited in environmental remediation techniques including heavy metal removal [6-8].

Bioemulsifier surfaced as promising remediation agent that can effectively remove metals from soil and water bodies. The Whole-genome sequence represents a valuable shortcut, helping
scientists find genes much more easily and quickly. It expected that being able to study the entire genome sequence will help in understanding how genes endeavour together to direct the maintenance, development, and growth of a whole organism. Besides, it can use to predict the genes involved in the synthesizing of biosurfactants in microbes[9-10]. The present study aimed to sequence the whole genome of biosurfactant producing *Bacillus tequilensis* using Next-Generation sequencing, De-novo assembly, genome annotation, variant calling and variant annotation.

2. Results and Discussion

2.1 Identification of Biosurfactant producing *Bacillus tequilensis*

Majority of biosurfactants are produced by the microbes such as *Pseudomonas* genus followed by *Bacillus* and *Acinetobacter* respectively [11]. In a previous investigation, a novel strain of biosurfactant producing *Bacillus tequilensis* strain ANSKLAB04 [12] was identified using 16S rRNA gene sequencing by Sanger dideoxy sequencing method followed by the phylogenetic assessment. The strain was isolated from Chilika lake, a brackish water lagoon, spread over the Puri, Khurda and Ganjam districts of Odisha state on the east coast of India [12]. By conducting several biochemical tests such as Haemolysis test, oil spreading test, CTAB agar plate test and Drop collapse test, we concluded that *Bacillus tequilensis* produces biosurfactants [12] and the novel isolate was deposited in Genbank with Accession number KU529483.

2.2 Bioanalyzer profile

The DNA isolation was performed using Phenol/Chloroform(PCI) genomic DNA extraction method[12]. The bioanalyzer profile of the prepared WGS library showed fragments in a size range of 300-600bp. The effective insert size of the library is 180-480bp flanked by adaptors having combined size of ~120bp. Based on the fragment distribution and concentration, the library was suitable for sequencing on Illumina platform [Fig 1].
2.3 Genome Representation

The complete genome of *Bacillus tequilensis* consists of a single circular chromosome of 4,478,749 bp with an average G+C content of 46.33% (Table 1 and Supplementary Table 1). The 4492 predicted coding ORFs covers 87% of the complete genome, and each ORF has a moderate length of 283 aa (Supplementary Table 2). Among these, 1,347, i.e. 67.4% assigned as putative functions, 258, i.e. 12.9% matched to sustain hypothetical coding sequences of an anonymous function, and the rest 394, i.e. 19.7% shows no similarities to known genes [Table 2].

Table 1: Assembly statistics of scaffolds:

| Assembly Stat            | Assembly |
|--------------------------|----------|
| Contigs Generated        | 528      |
| Maximum Contig Length    | 1664507  |
| Minimum Contig Length    | 500      |
|                          |       |
|--------------------------|-------|
| Average Contig Length    | 8482  |
| Median Contig Length     | 597   |
| Total Contigs Length     | 4478749 |
| Total Number of Non-ATGC Characters | 510 |
| Percentage of Non-ATGC Characters | 0.011 |
| Contigs >= 100 bp        | 528   |
| Contigs >= 200 bp        | 528   |
| Contigs >= 500 bp        | 528   |
| Contigs >= 1 Kbp         | 66    |
| Contigs >= 10 Kbp        | 12    |
| Contigs >= 1 Mbp         | 2     |
| N50 value                | 1077242 |

The variations in nucleotide frequencies across the whole genome sequence was investigated using a non-overlapping active platform and by framing three indices of nucleotide frequency: G+C%, (G+C)/(A+T+C+G), divergence from [A]=[T], (A-T)/(A+T), and divergence from [C]=(G), (C-G)/(C+G). These 3 indices are, by represent, pairwise-independent and summarize relative nucleotide frequencies without loss of information. Because of their very low frequency, ambiguous nucleotide bases were not taken into account. The SD (standard deviation) for the 3 indices are given by

\[
SD[(G + C)/(A + T + C + G)] = \frac{1}{N} \sqrt{\frac{SW}{N}}
\]  
(1)

\[
SD[(A - T)/(A + T)] = 2 \sqrt{\frac{AT}{W}}
\]  
(2)

\[
SD[(C - G)/(C + G)] = 2 \sqrt{\frac{CG}{S}}
\]  
(3)

Where, \(W = A + T\), \(S = C + G\) and \(N = A + T + C + G\).
Figure 2: **Genome Map of *Bacillus tequilensis***.

Normal distribution approximation was used as the total numbers of bases were large. The strand analyzed here was the 5' to 3' strand clockwise on the genetic map. A window size of 1 kb was used. From the inside: green and red bars represent RNA sequences on positive and negative strand respectively. Circle 1, represents G + C content (window size: 10Kb) higher and lower than 45%, where red represents higher and green represents lower. Circles 2: - represents GC skewness, where green and red represents positive and negative value respectively [Fig 2].
Table 2: Functional categories of predicted genes in *Bacillus tequilensis* genome

| COG categories                                      | No of Genes |
|-----------------------------------------------------|-------------|
| Information storage and processing                  |             |
| J. Translation, ribosomal structure and biogenesis   | 245         |
| A. RNA processing and modification                   | 25          |
| K. Transcription                                     | 231         |
| L. Replication, recombination and repair              | 238         |
| B. Chromatin structure and dynamics                   | 19          |
| Cellular Process                                     |             |
| D. Cell cycle control, cell division, chromosome partitioning | 72         |
| Y. Nuclear structure                                 | 2           |
| V. Defense mechanisms                                | 46          |
| T. Signal transduction mechanisms                    | 152         |
| M. Cell wall/membrane/envelope biogenesis            | 188         |
| N. Cell motility                                     | 96          |
| Z. Cytoskeleton                                      | 12          |
| W. Extracellular structures                          | 1           |
| U. Intracellular trafficking, secretion, and vesicular transport | 158     |
| O. Posttranslational modification, protein turnover, chaperones | 203     |
| Metabolism                                           |             |
| C. Energy production and conversion                  | 258         |
| G. Carbohydrate transport and metabolism             | 230         |
| E. Amino acid transport and metabolism               | 270         |
| F. Nucleotide transport and metabolism               | 95          |
| H. Coenzyme transport and metabolism                 | 179         |
| I. Lipid transport and metabolism                    | 94          |
| P. Inorganic ion transport and metabolism             | 212         |
| Q. Secondary metabolites biosynthesis, transport and catabolism | 88         |
| Poorly characterized                                 |             |
| R. General function prediction only                   | 702         |
| S. Function unknown                                  | 1347        |
| Not in COG                                           |             |

All genes were classified according to the COG classification. http://www.ncbi.nlm.nih.gov/COG/

2.4 Gene ontology and biological annotation

The gene ontology analysis concluded that 18.99% of genes in *Bacillus tequilensis* belonged to transferase activity, 13.55% of genes belonged to kinase activity, 9.3% of genes were involved in ATP binding, 9.3% genes were involved with hydrolase activity, 6.91% genes were involved in
methyltransferase activity, 5.98% of genes were associated with lipase activity, 5.98% of genes were involved in oxidoreductase activity, 4.9% of genes were in lyase activity, 3.05% genes were involved in peptidase activity, whereas only 2.79% genes were involved in cell division, 2.9% were in carbohydrate transport, 7.7% were in ribose production, and only 3.8% genes were involved in viral capsid [Fig 3].

Fig 3: Biological annotation of *Bacillus tequilensis*ANSKLAB04

### 2.5 Subsystem Classification

Genes obtained from the whole genome of *Bacillus tequilensis* has been used for the classification of the subsystem. Subsystems were categorized based on the cofactors, cell wall, virulence metabolism, potassium metabolism, membrane transport, iron acquisition and metabolism, RNA metabolism, cell division and cell cycle, motility and chemo taxis, fatty acids,
lipids and isoprenoids, nitrogen metabolism, etc were discussed in [Supplementary Table 3][Fig 4].

![Subsystem category distribution of Bacillus tequilensisANSKLAB04](image)

**Fig 4: Subsystem category distribution of Bacillus tequilensisANSKLAB04**

### 2.6 Metabolism of biosurfactant producing genes

*Bacillus tequilensis* produces a biosurfactant that belongs to the class of lipopeptides having excellent emulsifying properties and were capable of reducing the surface tension of water to a significantly lower value. The genes associated with producing biosurfactant are listed in [Table 3]. Among the several different classes of Biosurfactant producing bacteria genera, the members of the genera Bacillus or Pseudomonas, due to their wide range of applications and resourcefulness can be more often is used. Bacillus species are phenotypically and genotypically heterogeneous. Based on several investigations, a unique inhabitant of *Bacillus* sp. found at the marine site such as *B. subtilis, B. licheniformis, B. cereus, B. amyloliquefaciens, B. pumilus, and B. mycoides*. *Bacillus subtilis* produces lipopeptide biosurfactant called surfactin, which is coded by four ORFs named as SrfA, SrfB( also known as ComA), SrfC, and SrfD. The sfp gene considered as an essential component of peptide synthesis systems and plays a major role in the regulation of surfactinbiosynthesis and gene expression. Srf gene amplification is at 268 bp whereas the expression of the sfp gene amplified at 675 bp [13]. The peptide synthesise for an
The amino acid moiety of surfactin is encoded by four ORFs in the srfA operon namely SrfAA, SrfAB, SrfAC and SrfAD /SrfA-TE and also contains comS gene lying within the out-of-frame with the srfB [14]. PorobS et al 2013 and Nakano Met al., 1992 isolated SrfA gene from Bacillus amplified at 580 bp and authors concluded the biological significance of SrfA gene in biosurfactant production [15-17]. From Bacillus tequilensis we identified the SrfA which involved in biosurfactant production and the sequence of the SrfA(242 aa) was deposited in GenBank with accession MUG02427.1.

Besides, lichenysin is another lipopeptide biosurfactant produced by B. licheniformis coded by lichenysin operon (LchA) and comprises of four peptide synthetase genes: LicAA, LicAB, LicAC, and LicAD. In another study, the authors isolated genesfp (Phosphopantetheinyl transferase 224 amino acids) and mapped at 4kb downstream to operon srfA and the authors also concluded it is essential for the post-translational changes to surfactin synthetase in microbes [15 - 16]. In this study we have identified sfp gene from Bacillus tequilensis and the sequence of the sfp (Phosphopantetheinyl transferase 224 amino acids) was deposited in GenBank with accession MUG02422.1.

Moreover, two operons, srfA and pps were found to be present in UMX-103 and B. subtilis 168 strains only involved in biosurfactant synthesis. The srfA operon contains four genes such as srfAA, srfAB, srfAC, and srfAD and the operon pps contains four genes named as ppsB, ppsC, ppsD, and ppsE. The genes, rmlA, rmlB, rmlC and rmlD are only present in UMX -103 strains whereas, sigA, DnaK and LytR are present specifically in Bacillus strain. Besides, the genes comA, comP, rpoN, abrB and ResD are presented in both UMX-103 and B. subtilis 168 [18]. Based on the above literature biological annotation, we have identified DnaKandLytRgenes from Bacillus tequilensis and the sequence were deposited in NCBI with accession MUF99480.1 and MUG01692.1 respectively.

Pseudomonas species required Plasmid-encoded- rhlA, B, R and I genes of rhl quorum-sensing system for production of glycolipid biosurfactants as well as also involved in the production of rhamnolipids in a heterologous host. Iturin A is an antifungal lipopeptide biosurfactant produced by certain Bacillus subtilis strains such as B. subtilis RB14 is composed of four ORF namely
ituD, ituA, ituB, and ituC, whose disruption leads to specific deficiency in iturin A production. The three genes of arthrofactin operon of Pseudomonas namely arfA, arfB and arfC encode ArfA, ArfB and ArfC containing two, four and five functional modules respectively required for condensation, adenylation and thiolation. Besides, Amphisin is produced by Pseudomonas sp. DSS73 require gacS and amsY genes for the production of biosurfactant as these genes are mutants defective in the genes. Amphisin synthesis is regulated by the gacS gene as the gacS mutant regains the property of surface motility upon the introduction of a plasmid. Moreover, genes dnaK, dnaJ and grpE positively regulate the biosynthesis of putisolvin [14]. Putisolvin biosynthesis genes such as dnaK, dnaJ and grpE from Bacillus tequilensis were identified and the sequence were deposited in Genbank with accession MUF99480.1 MUF99481.1, MUF99479.1 respectively.

Acinetobacter species produces high molecular weight biosurfactants - Emulsan and Alasan with the involvement of gene. AlnA, AlnB and AlnC are essential for Alasan biosynthesis whereas wza, wzb, wzc, wzx, and wzy is required for Emulsan biosynthesis. For the production of fungal biosurfactant, emt1 and cyp1 are the two genes involved in the synthesis of these glycolipids and fb1 and hfb2 genes regulating the synthesis of hydrophobin [14]. Thus, gene plays a major role in the biosynthesis of various microbial surfactants, and hence the role of molecular genetics and gene regulation mechanisms in the production of biosurfactant is essential. In this study, we have identified biosurfactant producing genes and corresponding orfs of Bacillus tequilensis such gene SrfAD, SrfAC, SrfAA and the sequence of the same was deposited in Genbank database with accession MUG02427.1, MUG02428.1, MUG02429.1, MUG03515.1 respectively.

Table 3: Biosurfactants producing genes of Bacillus species

| S.No | Gene involved in Biosurfactant production | Reference |
|------|------------------------------------------|-----------|
| 1    | srf                                      | [13]      |
| 2    | sfp                                      | [13][14][15] |
| 3    | srfA                                     | [16]      |
| 4    | rhlB                                     | [16]      |
| 5    | cfp                                      | [17]      |
| 6    | srfAA                                    | [18]      |
| 7    | srfB                                     | [18]      |
| 8    | srfAB                                    | [18]      |
| 9    | srfAD                                    | [18]      |
|   |   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|---|
|10| Spf | [18] |
|11| ppsB | [18] |
|12| ppsC | [18] |
|13| ppsD | [18] |
|14| X ppsE | [18] |
|15| X dhbF | [18] |
|16| X rmlA | [18] |
|17| X rmlB | [18] |
|18| X rmlC | [18] |
|19| X rmlD | [18] |
|20| X comA | [18] |
|21| X comP | [18] |
|22| X ResD | [18] |
|23| X LiaR | [18] |
|24| spo0A | [18] |
|25| rpoN | [18] |
|26| X crsA | [18] |
|27| X sigA | [18] |
|28| abrB | [18] |
|29| X DnaK | [18] |
|30| LytR | [18] |

### 2.7 Biosurfactant / Lipopeptide metabolism of *Bacillus tequilensis*

Considering the biosurfactant producing genes described in various literatures, we classified the genes of *Bacillus tequilensis* based on the established efficient biosurfactant activity and broad applications. Biosurfactant is proven to be promising; possessing unique properties of low toxicity and higher biodegradability. In the present investigation we constructed a pathway which describes the biosurfactant metabolism of *Bacillus tequilensis*[Fig 5]. The lipopeptide synthesised constitutes of a long chain of fatty acid along with glutamate acid (Glu), leucine (Leu), aspartic acid (Asp) and valine (Val). The synthesis is non-ribosomal by a large multienzyme peptide, non-ribosome peptide synthases (NRPS). The peptide synthetase required for an amino acid moiety of surfactin is encoded by four open reading frames in the srfA operon namely SrfAA, SrfAB, SrfAC and SrfAD or SrfA-TE. SrfA, SrfB, SrfC and srfD constitute the four main enzymes for surfactin formation. SrfD is the most important enzyme as it initiates the surfactin formation. Sfp gene is 224 amino acid long- present downstream to srfA operon- also plays an important role as it is required for the posttranslational modifications to surfactinsynthetase.
Different modules have been marked based on different pathways involved for the synthesis of biosurfactant, such as glycolysis, TCA cycle, NADPH generation, amino acid biosynthesis, fatty acid synthesis and synthesis of surfactin. Seven modules represent the different pathways required for production of glutamate acid (Glu), leucine (Leu), aspartic acid (Asp) and valine (Val). The precursors for biosynthesis of Val/Leu, Glu/Asp, and fatty acids are the product of glycolysis and TCA cycle such aspyruvate, 2-oxo-glutarate, oxaloacetate, and acetyl-CoA. The genes of *Bacillus tequilensis* involved in the utilization of sucrose, including *sacP, murP*, and *sacA*, which encode a sugar transporter, permease, and sucrose-6-phosphate hydrolase, were identified and the sequence was deposited in genbank with accession MUF99868.1, MUG00557.1,
MUG01465.1 respectively. The NADPH generation and pentose are produced by pentose phosphate pathway catalysed by zwf and GNDA enzyme.

The biosynthesis of Glu, Asp, Val, and Leu, are considered as the intrinsic components of surfactin. Glu/Asp are synthesised by aspartate aminotransferase such as AspB and YhdR were identified from Bacillus tequilensis and the sequence were deposited to genbank using accession MUF99794.1 and MUF99877.1 respectively. The efficient fatty acid biosynthesis pathway determines efficient surfactin production. The building precursor acetyl-CoA initiates the biosynthesis of fatty acid. The biosynthesis of surfactin is catalysed through NRPS, initiated from the condensation of fatty acids and Glu. Other constituent amino acids are assembled through the NRPS multi-enzyme complex, comprising adenylation, condensation, and thiolation domains responsible for the activation of amino acids and peptide chain elongation.

2.8 Genome evolution of B. tequilensis

The enormous genomic data obtained from sequencing of Bacillus tequilensis ANSKLAB04 was aligned against existing top 20 homologous species of bacillus in the NCBI database. The comparison of the number of unique genes and common genes were analysed with top 6 homologous species of bacillus such as B. subtilis, B. vallismortis, B. tequilensis (KCTC 13622), B. halotoleran and B. mojavensis [Table 4] [Fig 6].

Table 4: Top 6 organisms homologous to B. tequilensis ANSKLAB04

| Sl | Organism | Accession | Size in mb | GC%  | Genes | Proteins | rRNA | tRNA | Pseudogenes |
|----|----------|-----------|------------|------|-------|----------|------|------|-------------|
| 01 | B. tequilensis ANSKLAB04 | RMVO01000000 | 4.38 | 46.33% | 4724 | 4492 | 28 | 81 | 118 |
| 02 | B. subtilis | AL009126.3 | 4.22 | 43.5% | 4536 | 4237 | 30 | 86 | 88 |
| 03 | B. vallismortis | CP026362.1 | 4.28 | 43.80% | 4514 | 4208 | 30 | 87 | 184 |
| 04 | B. tequilensis (KCTC 13622) | AYTO00000000.1 | 3.98 | 43.90% | 4167 | 3958 | 7 | 74 | 136 |
| 05 | B. halotoleran | CP029364.1 | 4.15 | 43.8% | 4298 | 4032 | 30 | 86 | 145 |
Fig 6: Comparison of Biosurfactant producing genes of *B. tequilensis* ANSKLAB04 with other species of Bacillus

The phylogenetic tree was constructed from the top 20 homologous bacillus species obtained from a blast search. The orange colors show the gene family expansion and grey color indicates the gene family contractions between the bacillus species. The corresponding proportions among the total changes are shown in the same colors in the pie chart. Implied divergence dates (in millions of years) are indicated at each node in blue. The most recent common ancestor (MRCA) and the blue color indicates the conserved gene family among the various species of Bacillus [Fig 7].
Fig 7: Phylogenetic affiliation of *Bacillus tequilensis*ANSKLAB04 against other existing species of bacillus

**2.9 SSR Identification**

Microsatellites or simple sequence repeats (SSRs) also known as short tandem repeats (STRs) are broadly used as PCR-based markers which can be helpful for characterization of population genetics, genetic mapping, phylogenetics, genome mapping, population genotyping, quantitative train loci analysis, genome comparisons, and markers assisted breeding etc. The present investigation predicted the accurate repeated SSRs with repeat unit length between 1 and 10 bp across the whole genome (528 scaffolds. i.e. 4478749 bp) of the whole genome of *Bacillus tequilensis*ANSKLAB04. K-mer distribution analysis found a total of 25 motif 2-mer, i.e. 86.2%, 3 motifs 3-mer i.e. 10.34% and motif 5 -mer i.e. 3.4% in the whole genome of *Bacillus tequilensis*ANSKLAB04 [Fig 8] [Fig 9].
Fig 8: K-mer distribution of motifs predicted from Bacillus tequilensis ANSKLAB04

Table 5: SSR patterns predicted from Bacillus tequilensis ANSKLAB04

| Scaffold_ID | Type of SSR | Pattern of SSR | SSR Length | Scaffold Start | Scaffold End | Scaffold length |
|-------------|-------------|----------------|-------------|----------------|-------------|----------------|
| scaffold1   | p1          | (A)10          | 10          | 1230382        | 1230391     | 1664507        |
| scaffold3   | p1          | (A)10          | 10          | 84253          | 84262       | 428649         |
| scaffold4   | p1          | (T)10          | 10          | 196300         | 196309      | 376322         |
| scaffold5   | p1          | (T)10          | 10          | 11858          | 11867       | 289205         |
| scaffold6   | p1          | (T)10          | 10          | 35289          | 35298       | 79869          |
| scaffold14  | p1          | (T)10          | 10          | 1371           | 1380        | 8363           |
| scaffold25  | p1          | (T)10          | 10          | 2110           | 2119        | 2203           |
| scaffold59  | p1          | (A)10          | 10          | 22             | 31          | 1087           |
| scaffold71  | p1          | (T)10          | 10          | 832            | 841         | 958            |
| scaffold239 | p1          | (A)10          | 10          | 544            | 553         | 614            |
| scaffold270 | p1          | (T)11          | 11          | 31             | 41          | 595            |
| scaffold280 | p1          | (A)13          | 13          | 453            | 465         | 591            |
| scaffold318 | p1          | (T)13          | 13          | 528            | 540         | 569            |
| scaffold328 | p1          | (G)22          | 22          | 541            | 562         | 564            |
| scaffold344 | p1          | (A)10          | 10          | 466            | 475         | 558            |
| scaffold2   | p2          | (AT)7          | 14          | 7978           | 7991        | 1077242        |
| scaffold72  | p2          | (TC)6          | 12          | 252            | 263         | 943            |
| scaffold85  | p2          | (TA)7          | 14          | 464            | 477         | 876            |
| scaffold150 | p2          | (TA)6          | 12          | 433            | 444         | 696            |
| scaffold317 | p2          | (TG)6          | 12          | 373            | 384         | 570            |
| scaffold455 | p2          | (TC)7          | 14          | 193            | 206         | 518            |
| scaffold34 | p3          | (TCA)5         | 15          | 923            | 937         | 1517           |
| scaffold95  | p3          | (ATG)5         | 15          | 620            | 634         | 810            |
| scaffold393 | p3          | (CCG)5         | 15          | 186            | 200         | 542            |
| scaffold491 | p5          | (GAATG)8       | 40          | 390            | 429         | 509            |
SSR composition was studied from each scaffold. The present investigation classified the SSR based on monomer, dimer, trimer, tetramer, pentamer, hexamer as P1, P2, P3, P4, P5, P6 respectively. The composition and repetition number of P1 to P6 lengthwise were studied and provided in [Table 5 - 6]. A total of 32 SSR were found from the 528 sequences of *Bacillus tequilensis* ANSKLAB04. 27 Sequences were identified with SSRs [Table 7]. A total of 27 P1 were found which recorded as the highest and P6 found the lowest [Fig 9].

**Table 6:** SSR types predicted from *Bacillus tequilensis* ANSKLAB04

| SSR Type                  | Set of Repeating Bases | Repetition number for the set | Example                  | Total Length       |
|---------------------------|------------------------|-------------------------------|--------------------------|--------------------|
| Mono nucleotide Repeats (p1) | 1                      | >= 10 bases                   | AAAAAAAAAAAAA           | >=10 to Any length |
| Di nucleotide Repeats (p2)    | 2                      | >= 6 Pairs                    | CACACACACACACA           | >=12 to Any length |
| Tri nucleotide Repeats (p3)    | 3                      | >= 5 Sets                     | ATGATGATGATG            | >=15 to Any length |
| Tetra nucleotide Repeats (p4)   | 4                      | >= 5 Sets                     | TGAGTGAGTGAGTGAG       | >=20 to Any length |
| Penta nucleotide Repeats (p5)   | 5                      | >= 5 Sets                     | TGAGCTGAGCTGAGCTGAG     | >=25 to Any length |
| Hexa nucleotide Repeats (p6)    | 6                      | >= 5 Sets                     | CATATACATACATACATACATACAATA | >=30 to Any length |

**Table 7:** Statistics of SSR Search predicted from *Bacillus tequilensis* ANSKLAB04

| Columns                                             |                        |
|-----------------------------------------------------|-------------------------|
| Total number of sequences examined:                 | 528                     |
| Total size of examined sequences (bp):              | 4478749                 |
| Total number of identified SSRs:                    | 32                      |
| Number of SSR containing sequences:                | 27                      |
| Number of sequences containing more than 1 SSR:     | 3                       |
| Number of compound SSRs (i.e c)                     | 4                       |
| p1                                                   | 21                      |
| p2                                                   | 7                       |
| p3                                                   | 3                       |
2.10 SNP and Indel Discovery

Our Indel discovery strategy involved mining insertion and deletion polymorphisms from DNA sequencing traces that originally were generated by genome centres for SNP discovery. The obtained mass sequenced data of *Bacillus tequilensis* ANSKLAB04 were used to search for genetic variation against existing homologous biosurfactant producing bacteria from NCBI. The present investigation used the existing to 5 homologous genomes of bacteria such as *Bacillus tequilensis* KCTC 13622, *Bacillus subtilis*, *Bacillus mojavensis*, *Bacillus vallismortis*, *Bacillus halotolerans*. The number of mapped sites per sample, mapping coverage, the total number of reads, number of mapped reads, overall mapping ratio, number of mapped bases, and the average alignment depth was calculated. Table 8 represents the statistics of *Bacillus tequilensis* on comparison with 5 existing homologous with reference bacterial genome which includes *Bacillus tequilensis* (KCTC 13622), *Bacillus halotoleran*, *Bacillus subtilis*, *Bacillus mojavensis* and *Bacillus vallismortis*. The number of total reads in all the reference genome is 6,229,938 which are constant in all reference bacteria. The mean depth indicates the number of reads, on average, are likely to be aligned at a given reference base position on comparison with *Bacillus tequilensis*. However, *Bacillus subtilis* is having 90.39% of mapped read, 786,017,247 mapped
bases and 186.45 mean depth which is highest among the other indicative of better analogy and susceptibility. On the other hand, *Bacillus mojavensis* with reference length 3,957,021, mapped reads 73.19%, mapped bases 555,891,216 and mean depth 140.48 showing the least compatibility with the *Bacillus tequilensis*.

Table 8: Mapped data Statistics of *Bacillus tequilensis* ANSKLAB04 against other homologous existing bacterial reference genome

| Ref Genome            | Ref Length | Mapped sites (>=1x) | Total reads | Mapped reads | Mapped bases | Mean Depth |
|-----------------------|------------|---------------------|-------------|--------------|--------------|------------|
| *Bacillus tequilensis* KCTC 13622 | 3,981,302  | 3,510,212 (88.17%)   | 6,229,938   | 5,236,833 (84.06%) | 702,871,686 | 176.54     |
| *Bacillus halotolerans* | 4,154,245  | 3,377,421 (81.3%)    | 6,229,938   | 4,720,660 (75.77%) | 576,944,088 | 138.88     |
| *Bacillus subtilis*    | 4,215,606  | 3,850,277 (91.33%)   | 6,229,938   | 5,631,246 (90.39%) | 786,017,247 | 186.45     |
| *Bacillus mojavensis*  | 3,957,021  | 3,251,746 (82.18%)   | 6,229,938   | 4,559,964 (73.19%) | 555,891,216 | 140.48     |
| *Bacillus vallismortis*| 4,286,362  | 3,466,929 (80.88%)   | 6,229,938   | 4,988,614 (80.07%) | 665,216,980 | 155.19     |

After removing duplicates with Sambamba and identifying variants with SAMTools, information of each variant were gathered and classified by chromosomes or scaffolds. Table 9 shows the summary of the variant calling of *Bacillus tequilensis* ANSKLAB04 against other existing genomes in the database.

Table 9 represents the summary of Variant Calling of *Bacillus tequilensis* against existing top 5 homologous references bacterial genome which includes *Bacillus tequilensis* (KCTC 13622), *Bacillus halotolerans*, *Bacillus subtilis*, *Bacillus mojavensis* and *Bacillus vallismortis*. Comparison of the whole-genome sequence of *Bacillus tequilensis* on comparison with a reference reveals the number of markers that include single nucleotide polymorphisms (SNPs), inserted and deleted sequences. Fig 10 represents the graphical representation of SNPs.
and INDEL in which *Bacillus halotolerans* is having the highest number of SNPs i.e. 347,175 [Figure 10D] whereas *Bacillus tequilensis* KCTC 13622 is having the maximum number of insertions and deletions i.e. 841 and 653 respectively [Figure 10A]. Meanwhile, *Bacillus subtilis* were having less number of SNPs, insertions and deletions [Figure 10B].

Table 9. Summary of Variant Calling of *Bacillus tequilensis* ANSKLAB04 against Existing species of Bacillus

| Ref Genome             | Library name                                      | Number of SNPs | Number of insertions | Number of deletions |
|------------------------|--------------------------------------------------|----------------|----------------------|---------------------|
| *Bacillus tequilensis* KCTC 13622 | SRR8203917(*Bacillus tequilensis ANSKLAB04*)       | 261,227        | 841                  | 653                 |
| *Bacillus subtilis*    | SRR8203917(*Bacillus tequilensis ANSKLAB04*)       | 47,864         | 496                  | 452                 |
| *Bacillus vallismortis*| SRR8203917(*Bacillus tequilensis ANSKLAB04*)       | 272,438        | 746                  | 604                 |
| *Bacillus halotolerans*| SRR8203917(*Bacillus tequilensis ANSKLAB04*)       | 347,175        | 671                  | 625                 |
| *Bacillus mojavensis*  | SRR8203917(*Bacillus tequilensis ANSKLAB04*)       | 338,879        | 692                  | 640                 |
Figure 10: SNP/Insertion/Deletion Count

2.10.1 Base Change Count

Table 10 represents the base change count on every SNPs of *Bacillus tequilensis* against Existing 5 existing homologous reference bacterial genome which includes *Bacillus tequilensis* (*KCTC 13622*), *Bacillus halotoleran*, *Bacillus subtilis*, *Bacillus mojavensis* and *Bacillus vallismortis* and [Fig 11A] are the graphical representation of base count change.
| Library Name | Ref | A   | C   |   |   |   |
|--------------|-----|-----|-----|---|---|---|
| Alt          | T   | G   | C   | A | T | G |
| SRR8203917   | 13,677 | 39,995 | 10,498 | 13,741 | 44,344 | 8,431 |
| Library Name | Ref |     |     |   |   |   |
| Alt          | A   | T   | C   | A | G | C |
| SRR8203917   | 44,784 | 13,579 | 8,575 | 13,462 | 10,271 | 39,870 |

| Library Name | Ref | A   | C   |   |   |   |
|--------------|-----|-----|-----|---|---|---|
| Alt          | T   | G   | C   | A | T | G |
| SRR8203917   | 2,391 | 8,082 | 2,021 | 1,997 | 8,227 | 1,142 |
| Library Name | Ref |     |     |   |   |   |
| Alt          | A   | T   | C   | A | G | C |
| SRR8203917   | 8,252 | 2,088 | 1,176 | 2,412 | 1,977 | 8,099 |

| Library Name | Ref | A   | C   |   |   |   |
|--------------|-----|-----|-----|---|---|---|
| Alt          | T   | G   | C   | A | T | G |
| SRR8203917   | 14,704 | 40,958 | 10,767 | 15,363 | 45,556 | 9,450 |
| Library Name | Ref |     |     |   |   |   |
| Alt          | A   | T   | C   | A | G | C |
| SRR8203917   | 45,463 | 14,764 | 9,510 | 14,616 | 10,525 | 40,762 |

| Library Name | Ref | A   | C   |   |   |   |
|--------------|-----|-----|-----|---|---|---|
| Alt          | T   | G   | C   | A | T | G |
| SRR8203917   | 18,913 | 51,771 | 17,789 | 17,062 | 53,943 | 14,151 |
| Library Name | Ref |     |     |   |   | T |
| SRR8203917   | |     |     |   |   | T |
|       | Alt | A   | T   | C   | A   | G   | C   |
|-------|-----|-----|-----|-----|-----|-----|-----|
| SRR8203917 |     | 53,584 | 16,973 | 14,202 | 19,145 | 17,655 | 51,987 |

**Bacillus tequilensis ANSKLAB04 vs. Bacillus mojavensis**

| Library name | Ref | A | C |
|--------------|-----|---|---|
| Alt          | T   | G | C | A | T | G |
| SRR8203917   | 18,917 | 51,594 | 17,012 | 16,395 | 51,610 | 13,798 |

| Library Name | Ref | G | T |
|--------------|-----|---|---|
| Alt          | A   | T | C | A | G | C |
| SRR8203917   | 51,821 | 16,592 | 13,723 | 18,824 | 17,310 | 51,282 |

Figure 11. Base change count of each sample
2.10.2 Transition and transversion information

The number of transition (Ts) and transversion (Tv), and the Ts/Tv ratio were calculated using the base change count. Base changes (DNA substitution) are of two types. Interchanges of purines (A <-> G), or pyrimidines (C <-> T) are transitions, while interchanges of a purine for pyrimidine bases, and vice versa, are transversions. Although there are twice as many possible transversions, transitions are more common than transversions due to differences in structural characteristics. Generally, transversions are more likely to cause amino acid sequence changes.

[Table 11 ] represents the transition and transversion information of Bacillus tequilensis against 5 existing homologous reference bacterial genome which includes Bacillus tequilensis(KCTC 13622), Bacillus halotoleran, Bacillus subtilis, Bacillus mojavensis, and Bacillus vallismortis and Figure [12] represents the proportional pie chart of Transversion and transition distribution. The transition/transversion ratio between homologous strands of DNA is generally about 2, but it is typically elevated in coding regions, where transversions are more likely to change the underlying amino acid and thus possibly lead to a fatal mutation in the translated protein. Bacillus halotolerans is having a maximum number of total SNPs counts hence their number of transition and transversion count is also more i.e. 211,285 and 135,890 respectively but the ratio percentage of Ts/Tv is 1.55 % is estimated by pairwise sequence comparison. On the other side, Bacillus subtilis is having the lowest count of total SNPs, Transition and Transversion but having the highest Ts/Tv ratio i.e 2.15%. Transition indicative number of A to T and C to G conversion or interchange and vice-versa whereas transversion is indicative of A to C or A to G or T to C or T to G or vice- versa as shown in figure [11].Bacillus subtilis is having more number of transitions on comparison with Bacillus tequilensis(Figure 12 B) i.e. 68.2 %. The number of transversions is more in Bacillus halotolerans and Bacillus mojavensis i.e. 39.1 %. However, in all 5 reference genome on comparison with Bacillus tequilensis, the count percentage of Transition is more than comparing to transversion (Figure 12 [D and E]). Transitions are less likely to result in amino acid substitutions and are therefore more likely to persist as "silent substitutions" in populations as single nucleotide polymorphisms (SNPs).
Table 11: Transition, Transversion information table

| Ref Genome             | Library Name                        | Total SNP Count | Transition | Transversion | Ts/Tv |
|------------------------|-------------------------------------|-----------------|------------|--------------|-------|
| *Bacillus tequilensis* | KCTC 13622 SRR8203917(Bacillus tequilensis ANSKLAB04) | 261,227         | 168,993    | 92,234       | 1.83% |
| *Bacillus subtilis*    | SRR8203917(Bacillus tequilensis ANSKLAB04) | 47,864          | 32,660     | 15,204       | 2.15% |
| *Bacillus vallismortis*| SRR8203917(Bacillus tequilensis ANSKLAB04) | 272,438         | 172,739    | 99,699       | 1.73% |
| *Bacillus halotolerans*| SRR8203917(Bacillus tequilensis ANSKLAB04) | 347,175         | 211,285    | 135,890      | 1.55% |
| *Bacillus mojavensis*  | SRR8203917(Bacillus tequilensis ANSKLAB04) | 338,879         | 206,308    | 132,571      | 1.56% |

Figure 12. Transition, Transversion proportion
2.11 Variant Annotation

To find out the annotation information such as amino acid changes by variants, SnpEff was used. Since genes usually have multiple transcripts, a single variant can have different effects on different transcripts. Table 8 and 9 shows the number of variants per type (based on the representative transcript), and brief explanations about the variant type, respectively. Top 10 types of variant annotations of *Bacillus tequilensis* ANSKLAB04 by comparing *Bacillus tequilensis* KCTC, *Bacillus subtilis*, *Bacillus vallismortis*, *Bacillus halotolerans*, *Bacillus mojavensis* [Table 12 - 16].

Table 12 represents the Annotation type count of *Bacillus tequilensis* ANSKLAB04 by comparing *Bacillus tequilensis* (KCTC 13622). There are various types of annotation found in *Bacillus tequilensis* (KCTC 13622) in comparison with *Bacillus tequilensis* ANSKLAB04. There upstream gene variant is having a maximum ratio of 98.27% with 256,198 indicative of a sequence variant located at 5' of a gene whereas the downstream gene variant indicative of a sequence variant located 3' of a gene which is 3,561(1.37%). Here is only 1 count of frameshift variant which indicates a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three is almost negligible.

**Table 12: Annotation type count of *Bacillus tequilensis* ANSKLAB04 by comparing *Bacillus tequilensis* KCTC 13622**

| Library name     | Type of annotation                      | Count   | Ratio  |
|------------------|-----------------------------------------|---------|--------|
| SRR8203917       | upstream_gene_variant                   | 256,198 | 98.27% |
|                  | downstream_gene_variant                 | 3,561   | 1.37%  |
|                  | intergenic_region                       | 780     | 0.3%   |
|                  | synonymous_variant                      | 128     | 0.05%  |
|                  | missense_variant                        | 40      | 0.02%  |
|                  | splice_region_variant                   | 5       | 0.0%   |
|                  | &non_coding_transcript_exon_variant     |         |        |
|                  | splice_region_variant                   | 5       | 0.0%   |
|                  | &stop_retained_variant                  |         |        |
|                  | disruptive_inframe_insertion            | 1       | 0.0%   |
|                  | initiator_codon_variant                 | 1       | 0.0%   |
Table 13 represents the annotation type count of *Bacillus tequilensis* ANSKLAB04 by comparing *Bacillus subtilis*. There are 7 types of annotations found in the comparison of *Bacillus subtilis* with *Bacillus tequilensis* ANSKLAB04 which includes upstream gene variant, downstream gene variant, intergenic region, synonymous variant, missense variant, initiator codon variant and disruptive inframe insertion. There upstream gene variant is having a maximum ratio of 96.92% with 47,287 indicative of a sequence variant located at 5' of a gene whereas downstream gene variant indicative of a sequence variant located 3' of a gene which is 1,098 (2.25%). Here synonymous variant count is 31 (0.06%) which is indicative of a sequence variant where there is no resulting change to the encoded amino acid.

**Table 13: Annotation type count of *Bacillus tequilensis* ANSKLAB04 by comparing *Bacillus subtilis***

| Library name | Type of annotation         | Count   | Ratio    |
|--------------|----------------------------|---------|----------|
| SRR8203917   | upstream_gene_variant      | 47,287  | 96.92%   |
|              | downstream_gene_variant    | 1,098   | 2.25%    |
|              | intergenic_region          | 363     | 0.74%    |
|              | synonymous_variant         | 31      | 0.06%    |
|              | missense_variant           | 10      | 0.02%    |
|              | initiator_codon_variant    | 2       | 0.0%     |
|              | disruptive_inframe_insertion| 1       | 0.0%     |

Table 14 represents Annotation type count of *Bacillus tequilensis* ANSKLAB04 by comparing *Bacillus vallismortis*. There are 9 types of annotation found in *Bacillus tequilensis* ANSKLAB04 in comparison with *Bacillus vallismortis*. There upstream gene variant is having the maximum ratio of 98.67% with 270,075 indicative of a sequence variant located at 5' of a gene whereas downstream gene variant indicative of a sequence variant located 3' of a gene which is 3,200 (1.17%).
Table 14: Annotation type count of *Bacillus tequilensis* ANSKLAB04 by comparing *Bacillus vallismortis*

| Library name          | Type of annotation                  | Count    | Ratio   |
|-----------------------|-------------------------------------|----------|---------|
| SRR8203917            | upstream_gene_variant               | 270,075  | 98.67%  |
|                       | downstream_gene_variant             | 3,200    | 1.17%   |
|                       | intergenic_region                   | 295      | 0.11%   |
|                       | synonymous_variant                  | 100      | 0.04%   |
|                       | missense_variant                    | 46       | 0.02%   |
|                       | splice_region_variant & stop_retained_variant | 6 | 0.0%   |
|                       | splice_region_variant & non_coding_transcript_exon_variant | 2 | 0.0%   |
|                       | disruptive_inframe_insertion        | 1        | 0.0%    |
|                       | initiator_codon_variant             | 1        | 0.0%    |

Table 15 represents Annotation type count of *Bacillus tequilensis* ANSKLAB04 by comparing *Bacillus halotolerans*. There are various types of annotation found in *Bacillus tequilensis* ANSKLAB04 in comparison with *Bacillus halotolerans*. There upstream gene variant is having the maximum ratio of 97.74% with 340,311 indicative of a sequence variant located at 5' of a gene whereas downstream gene variant indicative of a sequence variant located 3' of a gene which is 5,892 (1.69%).

Table 15: Annotation type count of *Bacillus tequilensis* ANSKLAB04 by comparing *Bacillus halotolerans*

| Library name          | Type of annotation                  | Count    | Ratio   |
|-----------------------|-------------------------------------|----------|---------|
| SRR8203917            | upstream_gene_variant               | 340,311  | 97.74%  |
|                       | downstream_gene_variant             | 5,892    | 1.69%   |
|                       | intergenic_region                   | 1,794    | 0.52%   |
|                       | synonymous_variant                  | 128      | 0.04%   |
|                       | missense_variant                    | 59       | 0.02%   |
|                       | splice_region_variant               | 10       | 0.0%    |
Table 16 represents Annotation type count of *Bacillus tequilensis ANSKLAB04* by comparing *Bacillus mojavensis*. There are various types of annotation found in *Bacillus tequilensis ANSKLAB04* in comparison with *Bacillus mojavensis*. There upstream gene variant is having the maximum ratio of 97.66% with 331,498 indicative of a sequence variant located at 5' of a gene whereas downstream gene variant indicative of a sequence variant located 3' of a gene which is 7,427 (2.19%).

**Table 16: Annotation type count of *Bacillus tequilensis ANSKLAB04* by comparing *Bacillus mojavensis***

| Library name | Type of annotation                                      | Count   | Ratio  |
|--------------|--------------------------------------------------------|---------|--------|
| SRR8203917   | upstream_gene_variant                                 | 331,498 | 97.66% |
|              | downstream_gene_variant                               | 7,427   | 2.19%  |
|              | intergenic_region                                     | 312     | 0.09%  |
|              | synonymous_variant                                    | 133     | 0.04%  |
|              | missense_variant                                      | 37      | 0.01%  |
|              | splice_region_variant&stop_retained_variant           | 11      | 0.0%   |
|              | splice_region_variant&non_coding_transcript_exon_variant | 6     | 0.0%   |
|              | initiator_codon_variant                               | 2       | 0.0%   |
|              | initiator_codon_variant&non_canonical_start_codon     | 1       | 0.0%   |

Variant calling tool SnpEff reports putative variant impact to make it easier and faster to categorize and prioritize variants. However, impact categories must be used with care as they were created only to help and simplify the filtering process. There is no way to predict whether a HIGH impact or a LOW impact variant is the one producing a phenotype of interest. The results
of the variant calling of *Bacillus tequilensis* ANSKLAB04 by comparing *Bacillus tequilensis KCTC 13622, Bacillus subtilis, Bacillus vallismortis, Bacillus halotolerans, Bacillus mojavensis* are provided in [supplementary material 4 – 8] respectively and annotation type information are provided in [Table 17].

Table 17: Annotation type information

| Type of annotation                        | Description                                                                 | Impact |
|-------------------------------------------|-----------------------------------------------------------------------------|--------|
| coding_sequence_variant                   | The variant hits a CDS.                                                     | MODIFIER |
| chromosome                               | A large part (over 1% or 1,000,000 bases) of the chromosome was deleted.     | HIGH   |
| duplication                               | Duplication of a large chromosome segment (over 1% or 1,000,000 bases).     | HIGH   |
| inversion                                | Inversion of a large chromosome segment (over 1% or 1,000,000 bases).       | HIGH   |
| coding_sequence_variant                   | One or many codons are changed.                                             | LOW    |
| disruptive_inframe_insertion              | One codon is changed and one or many codons are inserted (e.g.: An insert multiple of three at a codon boundary). | MODERATE |
| disruptive_inframe_insertion              | One codon is changed and one or many codons are inserted (e.g.: An insert multiple of three at a codon boundary). | MODERATE |
| inframe_deletion                          | One or many codons are deleted (e.g.: A deletion multiple of three at a codon boundary). | MODERATE |
| disruptive_inframe_insertion              | One codon is changed and one or more codons are deleted (e.g.: A deletion of size multiple of three, not at a codon boundary). | MODERATE |
| downstream_gene_variant                   | Downstream of a gene (default length: 5K bases).                           | MODIFIER |
| exon_variant                              | The variant hits an exon (from a non-coding transcript) or a retained intron.| MODIFIER |
| exon_loss_variant                         | A deletion removes the whole exon.                                          | HIGH   |
| exon_loss_variant                         | Deletion affecting part of an exon.                                         | HIGH   |
| duplication                               | Duplication of an exon.                                                     | HIGH   |
| duplication                               | Duplication affecting part of an exon.                                      | HIGH   |
| inversion                                | Inversion of an exon.                                                       | HIGH   |
| inversion                                | Duplication affecting part of an exon.                                      | HIGH   |
| frameshift_variant                        | Insertion or deletion causes a frame shift (e.g.: An indel size is not multiple of 3). | HIGH   |
| gene_variant                              | The variant hits a gene.                                                    | MODIFIER |
| feature_ablation                          | Deletion of a gene.                                                         | HIGH   |
| duplication                               | Duplication of a gene.                                                      | MODERATE |
| gene_fusion                               | Fusion of two genes.                                                        | HIGH   |
| gene_fusion                               | Fusion of one gene and an intergenic region.                                | HIGH   |
| bidirectional_gene_fusion                 | Fusion of two genes in opposite directions.                                 | HIGH   |
| rearranged_at_DNA_level                   | Rearrangement affecting one or more genes.                                 | HIGH   |
| intergenic_region                         | The variant is in an intergenic region.                                     | MODIFIER |
| Conserved_intergenic_variant              | The variant is in a highly conserved                                         | MODIFIER |
| Variant Type                  | Description                                                                 | Modifier |
|------------------------------|-----------------------------------------------------------------------------|----------|
| intergenic_variant           | The variant hits a gene, but no transcripts within the gene.                | MODIFIER|
| intron_variant               | Variant hits and intron. Technically, hits no exon in the transcript.       | MODIFIER|
| conserved_intron_variant     | The variant is in a highly conserved intronic region.                       | MODIFIER|
| miRNA                        | Variant affects an miRNA.                                                   | MODIFIER|
| missense_variant             | Variant causes a codon that produces a different amino acid (e.g.: Tgg/Cgg, W/R). | MODERATE|
| initiator_codon_variant      | Variant causes start codon to be mutated into another start codon (the new codon produces a different AA). (e.g.: Atg/Ctg, M/L (ATG and CTG can be START codons)) | LOW     |
| stop_retained_variant        | Variant causes stop codon to be mutated into another stop codon (the new codon produces a different AA). (e.g.: Atg/Ctg, M/L (ATG and CTG can be START codons)) | LOW     |
| protein_protein_contact      | Protein-Protein interaction loci.                                           | HIGH    |
| structural_interaction_variant| Within protein interaction loci (e.g. two AA that are in contact within the same protein, possibly helping structural conformation). | HIGH    |
| rare_amino_acid_variant      | The variant hits a rare amino acid thus is likely to produce protein loss of function. | HIGH    |
| splice_acceptor_variant      | The variant hits a splice acceptor site (defined as two bases before exon start, except for the first exon). | HIGH    |
| splice_donor_variant         | The variant hits a Splice donor site (defined as two bases after coding exon end, except for the last exon). | HIGH    |
| splice_region_variant        | A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron. | LOW     |
| splice_region_variant        | A variant affective putative (Lariat) branch point, located in the intron. | LOW     |
| splice_region_variant        | A variant affective putative (Lariat) branch point from U12 splicing machinery, located in the intron. | MODERATE|
| stop_lost                    | Variant causes stop codon to be mutated into a non-stop codon (e.g.: Tga/Cga, */R). | HIGH    |
| 5_prime_UTR_premature_start_codon_gain_variant | A variant in 5'UTR region produces a three base sequence that can be a START codon. | LOW     |
| start_lost                   | Variant causes start codon to be mutated into a non-start codon (e.g.: aTg/aGg, M/R). | HIGH    |
| stop_gained                  | Variant causes a STOP codon (e.g.: Cag/Tag, Q/*).                           | HIGH    |
| synonymous_variant           | Variant causes a codon that produces the same amino acid (e.g.: Ttg/Ctg, L/L). | LOW     |
| start_retained               | Variant causes start codon to be mutated into another start codon (e.g.: Ttg/Ctg, L/L (TTG and CTG can be START codons)). | LOW     |
| stop_retained_variant        | Variant causes stop codon to be mutated into another stop codon (e.g.: taA/taG, */*). | LOW     |
| transcript_variant           | The variant hits a transcript.                                              | MODIFIER|
| feature_ablation             | Deletion of a transcript.                                                   | HIGH    |
| regulatory_region_variant    | The variant hits a known regulatory region.                                 | MODIFIER|
| feature (non-coding) | description | MODIFIER |
|----------------------|-------------|----------|
| upstream_gene_variant| Upstream of a gene (default length: 5K bases). | MODIFIER |
| 3_prime_UTR_variant   | Variant hits 3'UTR region. | MODIFIER |
| 3_prime_UTR_truncation + exon_loss | The variant deletes an exon which is in the 3'UTR of the transcript. | MODERATE |
| 5_prime_UTR_variant   | Variant hits 5'UTR region. | MODIFIER |
| 5_prime_UTR_truncation + exon_loss_variant | The variant deletes an exon which is in the 5'UTR of the transcript. | MODERATE |

**Description for the table 17**

Type of annotation: Sequence ontology which allows to standardize terminology used for assessing sequence changes and impact.

Description: Detailed description of the effect (annotation).

Impact: Effects are categorized by 'impact': {High, Moderate, Low, Modifier}. These are pre-defined categories to help users find more significant variants.

HIGH: The variant is assumed to have high (disruptive) impact on the protein, probably causing protein truncation, loss of function or triggering nonsense mediated decay.

MODERATE: A non-disruptive variant that might change protein effectiveness.

LOW: Assumed to be mostly harmless or unlikely to change protein behavior.

MODIFIER: Usually non-coding variants or variants affecting non-coding genes, where predictions are difficult or there is no evidence of impact.

**Description for the supplementary tables 4-8**

Chromosome: Chromosome name.

Pos: Position information of target variant.

Ref: Reference sequence regarding specific position.

Alt: DNA sequence of the sample.

Quality: Phred-scaled probability of all samples being homozygous reference. The value is in -log. The smaller the value, the more likely ALT is wrong.

Hom/Het: Indicates the genotype. "hom" refers to non-reference homozygote, while "het" refers to heterozygote.

- Homozygous: The circumstances when there are mutations on most reads that are mapped to a certain region.
- Heterozygous: The circumstances when there is mutations on some reads that are mapped to certain region.
Read Depth: Total read depth.
Alt Depth: Allelic depths for the ref and alt alleles in the order listed.
Gene Name, GeneID: Gene name and gene symbol.
Start, End: Position information of target gene.
Strand: Strand information of target gene.
Transcript: The results of functional annotation by transcripts. Type of variant (syn/nonsyn), protein change, and etc. can be ascertained in this section. A representative transcript is chosen by the gene name obtained from variant calling analysis. Other transcripts are chosen by information of neighboring genes which are close enough.
- It is not uncommon for a gene to have more than one transcript. A variant might affect different transcripts in different ways, as a result of different reading frames.

3. DISCUSSION
In the present investigation, we have introduced a high-quality draft genome sequence of *Bacillus tequilensis*, the first genome sequence of biosurfactant producing *Bacillus tequilensis* has been determined. Biosurfactant producing microbes have potential applications in various biotechnology, biodegradation, pharmaceutical industries. The whole genome sequence of biosurfactant producing *Bacillus tequilensis* will provide a foremost resource to start exploring the genes and gene products involved in biosurfactant synthesis. The genome sequence of *Bacillus tequilensis* obtained in the present investigation will be a key resource for the development of new concept and technique in genetic engineering such as molecular marker assisted breeding and large scale production of biosurfactant microbes for bioremediation.

URLs.
FASTQC: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
Trimmomatic: http://www.usadellab.org/cms/?page=trimmomatic
SPAdes Assembler: https://github.com/ablab/spades

SSPACE: http://www.baseclear.com/bioinformatics-tools/

NCBI Prokaryotic Genome Annotation Pipeline:
https://www.ncbi.nlm.nih.gov/genome/annotation_prok/

KAAS - KEGG Automatic Annotation Server: https://www.genome.jp/tools/kaas/

COG: https://www.ncbi.nlm.nih.gov/COG/

MISA: https://webblast.ipk-gatersleben.de/misa/

BWA: http://bio-bwa.sourceforge.net/

Sambamba: http://lomereiter.github.io/sambamba/

SAM tools: http://samtools.sourceforge.net/

SnpEff: https://pcingola.github.io/SnpEff/

4. Methods

4.1 Sample Collection and DNA Isolation

Water samples were collected from oil-contaminated sites of Chilika lake, Odisha, India (latitude and longitude: 19.8450 N 85.4788 E), a largest brackish water lagoon in India. Various organisms were isolated and purified on culture plates and were then enriched in the mineral salt medium (MSM). MSM gives the nutrient condition for the production of biosurfactants by the organisms which were then screened for their biosurfactant production by various screening tests and the emulsification index was calculated. Identification of organisms was performed based on biochemical, macroscopic and microscopic characters. The organism with the best emulsification index was then subjected to optimization for the production of biosurfactant for the factors affecting the production. Optimization was studied with the emulsification index calculated with each affecting factor. In a previous study, this organism was then subjected to 16S rRNA sequencing for the identification of the genus and species [12]. The DNA was isolated by
Phenol/Chloroform (PCI) genomic DNA extraction method [12][19]. The bacterial cell pellet obtained after centrifugation was subjected to DNA isolation. The DNA concentration and purity were checked with nanodrop spectrophotometer and qubit fluorometer. ImageQuant software was used to analyse the gel cropped image documentation [20]. The sample was found to have suboptimal concentration and gave an intact band when running on gel QC against Hind III digested lambda ladder. The gel image demonstrates the intact *Bacillus tequilensis*. Whole genome having a concentration of 54.9 ng/ul. L represents Hind III Digested lambda ladder. '1 - SO_4915_Bt1' represents the *Bacillus tequilensis* sample code. [Fig 13].

Figure 13: DNA concentration and purity of samples estimated using Nanodrop Spectrophotometer and Qubit Flurometer

4.2 Materials used in the study

Whole Genome Sequencing kits such as NEXTFlex DNA Sequencing Kit (Cat # 5140-02), NEXTFlex DNA Barcodes – 48 (Cat # 514104), HighPrep™ PCR (Magbio, #AC-60050), High Sensitivity Bioanalyzer Chips (Agilent, #5067-4626), Nuclease free water (Ambion, #AM9939), Covaris™ S220 System (Life Technologies, #4465653), Covaris™ microTUBE AFA (Life Technologies, #520045), Low Melting Agarose (Invitrogen, #16520100), MinElute Gel Extraction kit (QIAGEN, #28604), 50X TAE Buffer (MP Biomedical, Cat #TAE50X01), Qubit® dsDNA HS Assay Kit (Invitrogen, Cat # Q32854) were used in the present investigation.
4.3 Library Preparation and Genome Sequencing

Library preparation was performed using NEXTFlex DNA library protocol outlined in “NEXTFlex” DNA sample preparation guide (Cat # 5140-02). In brief, genomic DNA was sheared to generate fragments of approximately 300-500bp in a CovarismicroTube with the E220 system (Covaris, Inc., Woburn, MA, USA). The fragment size distribution was checked using Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA) with High Sensitivity DNA Kit (Agilent Technologies) according to the manufacturer’s instructions. The resulting fragmented DNA was cleaned up using HighPrep beads (MagBio Genomics, Inc, Gaithersburg, Maryland). These fragments were subjected to end-repair, A-tailing, and ligation of the Illumina multiplexing adaptors using the NEXTFlex DNA Sequencing kit as per the manufacturer’s instruction [21].
The resulting ligated DNA was cleaned up using HighPrep beads (MagBio Genomics, Inc, Gaithersburg, Maryland) and size selected (400–600bp) on 2% low melting agarose gel and cleaned using MinElute column (QIAGEN, India). These adapter-ligated fragments were subjected to 10 rounds of PCR (denaturation at 98°C for 2 min, cycling (98°C for the 30S, 65°C for 30S and 72°C for 1 min) and a final extension at 72°C for 5 min) using primers provided in the NEXTFlex DNA Sequencing kit (Perkin Elmer). The PCR products were purified using HighPrep beads. Quantification and size distribution of the prepared library was determined using Qubit fluorometer (Table 18) and the Agilent High Sensitivity DNA Kit (Agilent Technologies) respectively according to the manufacturer’s instructions (Fig 14). Illumina
Paired-end sequencing was performed using NextSeq 500: 150*2. The following adapters were used for sequencing (Illumina, Inc)[21].

Adapter details: Universal Adapter 5’
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT 3’
Adapter, Index 5’
GATCGGAAGAGCACACGTCTGAACTCCAGTCAC[INDEX]ATCTCGTATGCGCTCTTCCTGCTGTG 3’

Table 18: Library concentration estimation using Qubit

| Sample ID          | Qubit Conc. (ng/μl) | Vol (l) | Yield(ng) | Nextflex Barcode | Barcode Sequence | qPCR conc. (nM) |
|--------------------|---------------------|---------|-----------|------------------|-----------------|-----------------|
| SO_4915_Bt1_ePCR1_IL_WGS | 3.92                | 12      | 47.04     | 4                | GCCAAT          | 7.092           |

4.3 Whole Genome De-novo Assembly and Analysis

The obtained sequence raw reads were checked for quality control using FASTQC tool [22]. The quality of the raw reads was checked through the various modules provided by the FASTQC tool. Among the modules, per base sequence quality and tile sequence quality modules were studies to validate the quality of the data for further analysis. The low-quality reads were excluded from the analysis using Trimmomatic (v0.36)[23]. The filtered De-novo assembly of Illumina paired-end data was assembled using SPAdes - v3.13.0 genome assembler - an open-source algorithm for De-novo assembly [24]. SPAdes assembler is intended for de-novo assembly after error-correction of sequenced reads. Assembled contigs were further scaffolded using SSPACE program [25]. Genome map was constructed using Circos [26].

4.4 Whole Genome annotation and GO analysis

NCBI Prokaryotic Genome Annotation Pipeline (PGAP) version 4.8 was used to annotate the whole genome sequence of [27]. Pathway Analysis was done by using KAAS Server. *Bacillus subtilis subsp. Subtilis 168* was taken as a reference organism for pathway analysis using KAAS server [28]. The functions of the predicted ORFs were categorized by comparison with the COG
database [29]. Simple Sequence Repeats (SSR) was identified in each transcript sequence using MISA Perl script [30].

4.5 Variant Calling and Variant annotation

Variant Calling of *Bacillus tequilensis* was performed by aligning with the top 5 existing homologous reference bacterial genome which includes *Bacillus tequilensis* (KCTC 13622), *Bacillus halotoleran*, *Bacillus subtilis*, *Bacillus mojavensis* and *Bacillus vallismortis*. The present investigation used BWA (Burrows-Wheeler Aligner)-MEM for alignment of *Bacillus tequilensis* against top 5 homologous genomes [31]. During mapping, duplicated reads can falsely cause erroneous data to stand out. To prevent such error, Sambamba tool was used to remove the duplicate reads [31]. Duplicate reads are identified using mapping information such as start position, and CIGAR string [32]. SAMTools was used to manipulate the SAM/BAM files that come out as a result of mapping [33]. In resequencing analysis, it is especially used for finding out variant information by calculating genotype likelihood from every position within the sample of analysis. Variant annotation was performed using SnpEff (v4.3t) [34]. SnpEff annotates the possible effects (on genes) that can be caused by variants identified through mapping. The present study used SnpEff to generate the Genes and transcripts affected by the variant, Location of the variants and the information on how the variant affects the protein synthesis (e.g. generating a stop codon).
Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials.
This whole-genome shotgun project has been deposited in GenBank/ENA/DDBJ under the Accession: RMVO00000000. The short-read sequences have been deposited under BioProject Accession: PRJNA498807, BioSample Accession: SAMN10335300 and SRA Accession: SRX5023292.

WGS URL: https://www.ncbi.nlm.nih.gov/nuccore/RMVO00000000
Bioproject URL: https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA498807
Biosample URL: https://www.ncbi.nlm.nih.gov/biosample/SAMN10335300
SRA URL: https://www.ncbi.nlm.nih.gov/sra/?term=SRX5023292

Competing interests
The authors declare no potential conflicts of interests.

Funding
Not applicable.

AUTHOR CONTRIBUTIONS
AN and SKS designed and coordinated the project. AN Collected the samples, cultured the bacteria and isolated the genomic DNA. AN and SKS sequenced and processed the raw data, assembled the genome, annotated the whole genome, analyzed the gene families, conducted
genome evolution analysis, conducted the variant calling, variant annotation and drafted the manuscript.

ACKNOWLEDGMENTS

The authors are thankful to Eminent Biosciences and LeGene Biosciences Pvt Ltd, Indore, India for 16S rRNA sequencing and Whole Genome Sequencing and De novo assembly of the bacterium.

References

1. Mulligan, C. N. (2005). *Environmental pollution*, 133(2), 183-198.
2. Banat, I. M. (1995). *Bioresource technology*, 51(1), 1-12.
3. Lin, S. C. (1996). *Journal of Chemical Technology and Biotechnology*, 66(2), 109-120.
4. Sheng, X.; Xia, J.J. (2006). *Chemosphere* 64, 1036-1042.
5. Sheng, X.; He, L.; Wang, Q.; Ye, H.; Jiang, C. (2008). *J. Hazard. Mater.* 155(1-2), 17-22.
6. Mulligan, C. (2005). *Environ. Pollut.* 133, 183-198.
7. Franzetti, A.; Caredda, P.; Ruggeri, C.; La Colla, P.; Tamburini, E.; Papacchini, M.; Bestetti, G. (2009). *Chemosphere* 75(6), 801-807.
8. Dahrazma, B.; Mulligan, C.N. (2007). *Chemosphere* 69(5),705-711.
9. Stein, L. (2001). *Nature Reviews Genetics* 2(7): Pages 493–503.
10. Wade, Nicholas (2007-05-31). The New York Times. Retrieved 2010, Pages 04-02.
11. Khire, J. M. (2010). Bacterial biosurfactants, and their role in microbial enhanced oil recovery (MEOR). In *Biosurfactants* (pp. 146-157). Springer, New York, NY.
12. Nayarisseri, A., Singh, P., & Singh, S. K. (2019). Screening, isolation and characterization of biosurfactant-producing *Bacillus tequilensis* strain ANSKLAB04 from brackish river water.International Journal of Environmental Science and Technology, Springer nature. 16(11), 7103-7112.
13. Swaathy, S., Kavitha, V., Sahaya Pravin, A., Sekaran, G., Mandal, A. B., &Gnanamani, A. (2014). Phylogenetic framework and biosurfactant gene expression analysis of marine Bacillus spp. of Eastern Coastal Plain of Tamil Nadu. *International journal of bacteriology, 2014*.
14. Das, P., Mukherjee, S., & Sen, R. (2008). Genetic regulations of the biosynthesis of microbial surfactants: an overview. *Biotechnology and Genetic Engineering Reviews*, 25(1), 165-186.
15. Porob, S., Nayak, S., Fernandes, A., Padmanabhan, P., Patil, B. A., Meena, R. M., & Ramaiah, N. (2013). PCR screening for the surfactin (sfp) gene in marine Bacillus strains and its molecular characterization from *Bacillus tequilensis* NIOS11. *Turkish Journal of Biology*, 37(2), 212-221.
16. Nakano, M. M., Corbell, N., Besson, J., & Zuber, P. (1992). Isolation and characterization of sfp: a gene that functions in the production of the lipopeptide biosurfactant, surfactin, in Bacillus subtilis. Molecular and General Genetics MGG, 232(2), 313-321.

17. Sekhon, K. K., Khanna, S., & Cameotra, S. S. (2011). Enhanced biosurfactant production through cloning of three genes and role of esterase in biosurfactant release. Microbial cell factories, 10(1), 49.

18. Abdelhafiz, Y. A., Manaharan, T., BinMohamad, S., & Merican, A. F. (2017). Draft Genome Sequence of a Biosurfactant-Producing Bacillus subtilis UMX-103 Isolated from Hydrocarbon-Contaminated Soil in Terengganu, Malaysia. Current microbiology, 74(7), 803-805.

19. Bonifer, K.S., Wen, X., Hasim, S., Phillips, E.K., Dunlap, R.N., Gann, E.R., DeBruyn, J.M. and Reynolds, T.B. (2019). Bacillus pumilus B12 Degrades Polylactic Acid and Degradation Is Affected by Changing Nutrient Conditions. Frontiers in Microbiology, 10, 2548.

20. Wheelock, Å. M., & Buckpitt, A. R. (2005). Software-induced variance in two-dimensional gel electrophoresis image analysis. Electrophoresis, 26(23), 4508-4520.

21. Steemers, F. J., & Gunderson1, 2, K. L. (2005). Illumina, Inc.

22. Andrews, S. (2010). Babraham bioinformatics-FastQC a quality control tool for high throughput sequence data. URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc.

23. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170.

24. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology. 2012;19: 455-477.

25. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., & Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics, 27(4), 578-579.

26. Krzywinski, Martin, et al. "Circos: an information aesthetic for comparative genomics." Genome research 19.9 (2009): 1639-1645.

27. Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E.P., Zaslavsky, L., Lomsadze, A., Pruitt, K.D., Borodovsky, M. and Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline. Nucleic acids research, 44(14), 6614-6624.

28. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., & Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic acids research, 35(suppl_2), W182-W185.

29. Tatusov, R. L., Galperin, M. Y., Natale, D. A., & Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic acids research, 28(1), 33-36.

30. Beier, S., Thiel, T., Münch, T., Scholz, U., & Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. Bioinformatics, 33(16), 2583-2585.
31. Abuín, J. M., Pichel, J. C., Pena, T. F., & Amigo, J. (2015). BigBWA: approaching the Burrows–Wheeler aligner to Big Data technologies. Bioinformatics, 31(24), 4003-4005.
32. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., & Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. Bioinformatics, 31(12), 2032-2034.
33. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078-2079.
34. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin) 6: 80–92.