Data Article

Data on genome sequencing, analysis and annotation of a pathogenic *Bacillus cereus* 062011msu

Rashmi Rathya¹, Sayan Paula¹, Vasanthakumar Ponesakki¹, Paulkumar Kanniah², Suriya Prabha Muthu³, Arun Arumugaperumal¹, Emmanuel Joshua Jebasingh Sathiya Balasingh Thangapandi¹, Subburathan Balakrishnan¹, Rajendran Jeyaprakash², Sudhakar Sivasubramaniam¹,²

¹ Department of Biotechnology, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu 627012, India
² Department of Genetics, School of Biological Sciences, Madurai Kamaraj University, Madurai, Tamil Nadu 625021, India

**Abstract**

*Bacillus* species 062011 msu is a harmful pathogenic strain responsible for causing abscessation in sheep and goat population studied by Mariappan et al. (2012) [1]. The organism specifically targets the female sheep and goat population and results in the reduction of milk and meat production. In the present study, we have performed the whole genome sequencing of the pathogenic isolate using the Ion Torrent sequencing platform and generated 458,944 raw reads with an average length of 198.2 bp. The genome sequence was assembled, annotated and analysed for the genetic islands, metabolic pathways, orthologous groups, virulence factors and antibiotic resistance genes associated with the pathogen. Simultaneously the 16S rRNA sequencing study and genome sequence comparison data confirmed that the strain belongs to the species *Bacillus cereus* and exhibits 99% sequence homology with the genomes of *B. cereus* ATCC 10987 and *B. cereus* FRI-35. Hence, we have renamed the organism as *Bacillus cereus* 062011msu. The Whole Genome Shotgun (WGS) project has

---

**Article Info**

**Article history:**
Received 1 November 2017
Received in revised form 16 December 2017
Accepted 28 December 2017
Available online 3 January 2018

**Keywords:**
*Bacillus cereus*
Genome sequencing
Abscession
Virulence factors

**Abstract**

*Bacillus* species 062011 msu is a harmful pathogenic strain responsible for causing abscessation in sheep and goat population studied by Mariappan et al. (2012) [1]. The organism specifically targets the female sheep and goat population and results in the reduction of milk and meat production. In the present study, we have performed the whole genome sequencing of the pathogenic isolate using the Ion Torrent sequencing platform and generated 458,944 raw reads with an average length of 198.2 bp. The genome sequence was assembled, annotated and analysed for the genetic islands, metabolic pathways, orthologous groups, virulence factors and antibiotic resistance genes associated with the pathogen. Simultaneously the 16S rRNA sequencing study and genome sequence comparison data confirmed that the strain belongs to the species *Bacillus cereus* and exhibits 99% sequence homology with the genomes of *B. cereus* ATCC 10987 and *B. cereus* FRI-35. Hence, we have renamed the organism as *Bacillus cereus* 062011msu. The Whole Genome Shotgun (WGS) project has
### Specifications Table

| Subject area | Biology |
|--------------|---------|
| More specific subject area | Bioinformatics (Genomics) |
| Type of data | Table, figure |
| How data was acquired | Genome sequencing: Ion Torrent personal genome machine (PGM) (Life Technologies, Carlsbad, CA), Denovo sequence assembly: CLC genomics workbench version 9.0.1, Bioinformatics approaches: NCBI Prokaryotic Genomes Automatic Annotation Pipeline (PGAAP), RAST genome annotation server (http://rast.nmpdr.org/), EggNog database integrated in BLAST2GO (eggno.g.embl.de; https://www.blast2go.com/), MEGA7 (Multiple sequence alignment and phylogenetic analysis), DNAPlotter (http://www.sanger.ac.uk/science/tools/dnaplotter). |
| Data format | Analyzed |
| Experimental factors | Genome sequencing, genome annotation, KEGG pathway analysis, orthologous group analysis, 16S rRNA and 23S rRNA based phylogeny. |
| Experimental features | The whole genome sequencing of *Bacillus cereus* 062011msu was performed by using Ion Torrent Personal Genome Machine (PGM) platform. Quality analysis, filtering and de novo assembly of the raw reads were performed by CLC genomics workbench 9.0.1. Genome annotation was done by using the PGAAP pipeline and RAST genome annotation server. Multiple sequence alignment and phylogenetic analysis based on 16S rRNA and 23S rRNA sequences were performed by the MEGA7 tool. The circular genome map of our species was generated by DNAPlotter. |
| Data source location | Maruthamputhur village, Alangulam, Tirunelveli District, Tamil Nadu, India. (latitude: 8.8646N and longitude: 77.4960 E). |
| Data accessibility | Genome analysis and annotation data are given within this article and the raw data along with NCBI PGAAP annotation were deposited at NCBI repository: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA404036, Bioproject ID: 404036, BioSample: SAMN07629099. The Whole Genome Shotgun (WGS) project has been deposited at DDBJ/ENA/GenBank under the accession NTMF000000000 (https://www.ncbi.nlm.nih.gov/nuccore/NTMF000000000). The genome annotation data obtained from the RAST server are given in this article. |
| Related research article | "Bacillus sp. causing absccessation in sheep and goat population" by Mariappan et al. (2012) [1]. |
Value of the data

- The *Bacillus cereus* 062011msu is a deadly pathogenic bacterium known for causing abscess mainly in the female sheep and goat population. Hence, the genome sequence resource and their annotation details can be effectively utilized to understand the pathogenicity of the bacterium for the benefit of the farmers who rear the sheep and goat.
- The genome annotation data of *Bacillus cereus* 062011msu provided a broad overview regarding the subsystem features, metabolic pathways, orthologous groups, virulence factors and antibiotic resistant genes associated with the genome of the species. Most of the unique genes of the species were found to be clustered in ten genetic islands. In this study we provided a detailed analysis of the genes clustered on the genetic islands.
- The data obtained from 16S rRNA analysis and genome sequence comparison with other *Bacillus* species provided significant information regarding the identification and taxonomic classification of this new bacterial strain. Although according to the previous study using the partial 16S RNA sequences the pathogen was reported to be genetically similar to *Bacillus anthracis* [1], but the whole genome data confirmed that the strain is in fact belongs to the species *Bacillus cereus* and phylogenetically related with *B. cereus* ATCC 10987 and *B. cereus* FRI-35.
- The entire genome dataset can be utilized further for determining the genes and biochemical pathways related to the pathogenicity (abscess) of the strain and developing new antimicrobial drugs for the pathogen.

1. Data

The overall data represents the genome sequencing, assembly, annotation and comparative analysis of pathogenic bacteria *Bacillus cereus* 062011msu. Table 1 denotes the summary statistics of the draft genome assembly of the *B. cereus* 062011msu. The data describing the length and Phred quality score distribution of the raw and filtered reads are illustrated in Supplementary Fig. S1. Data on Fig. 1 represent 10 genetic islands predicted in the genome of the isolate. The details of the genes clustered on the genetic islands are shown in Supplementary Table S1. Fig. 2 shows the subsystem distribution of the *B. cereus* 062011msu genome based on RAST genome annotation. The complete list of the RAST annotated genes is given in Supplementary Table S2. Fig. 3 gives a complete overview of the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways associated with the annotated genome sequence. The data illustrated in Fig. 4 show the Clusters of Orthologous Groups (COG) distribution of

Table 1
Summary statistics of *Bacillus cereus* 062011msu genome assembly.

| No. of raw reads | 458,944 |
|------------------|---------|
| average quality PHRED-score Q20 (before trimming) | 26,143 (abs) |
| average quality PHRED-score Q25 (before trimming) | 45,447 (abs) |
| average quality PHRED-score Q30 (before trimming) | 21,587 (abs) |
| No. of clean reads | 432,619 |
| Average length of clean reads | 178.08 |
| average quality PHRED-score Q20 (after trimming) | 5096 (abs) |
| average quality PHRED-score Q25 (after trimming) | 54,520 (abs) |
| average quality PHRED-score Q30 (after trimming) | 29,972 (abs) |
| Percentage of clean reads | 94.27% |
| Total no. of assembled contigs | 3200 |
| Mean contig length | 1611 |
| N50 length | 2922 |
| N25 length | 5365 |
| Sequencing depth | 15X |
| GC% | 35.3% |
| Size of genome | 5,154,790 |

Abs: number of sequences observed at that quality score.
the protein coding genes obtained from RAST annotation. Table 2 denotes the list of virulence factors (with complete homology) identified in the annotated genome dataset. The data regarding the antibiotic resistance genes identified in the pathogen are given in Supplementary Table S3. The top 20 closest neighboring strains of Bacillus cereus 062011msu based on RAST annotation are listed in Supplementary Table S4. Supplementary Table S5 portrays the top 10 species showing maximum sequence homology with the genome of Bacillus cereus 062011msu resulted from BLAST genome alignment. Fig. 5 represents the phylogenetic tree constructed based on 16S rRNA comparison of the strain with its closely related homologs. Simultaneously the phylogenetic analysis data obtained from 23S rRNA comparison study are depicted in Fig. S2. Data on Fig. 6 represents the visualization of the annotated circular genome map of B. cereus 062011msu obtained from DNAPlotter.
2. Experimental design, materials and methods

2.1. Genome sequencing, quality assessment and de novo assembly

The *Bacillus cereus* 062011msu was isolated from the abscess tissue of the affected female sheep and goats in Maruthampthur village near Alangulam Region, Tirunelveli District, Tamil Nadu, India [1]. The whole genome sequencing of the species using Ion Torrent personal genome machine (Life Technologies, Carlsbad, CA) [2] produced 458,944 raw reads having average length of 198.2 bp and

**Fig. 3.** Pie chart representing the distribution of KEGG pathways associated with the genome of *Bacillus cereus* 062011msu. The pathways were obtained by annotating the protein coding sequences against the KEGG database.

**Fig. 4.** Clusters of Orthologous Groups (COG) distribution of the protein coding genes in *Bacillus cereus* 062011msu obtained from EggNog database.

2.1. Genome sequencing, quality assessment and de novo assembly

The *Bacillus cereus* 062011msu was isolated from the abscess tissue of the affected female sheep and goats in Maruthampthur village near Alangulam Region, Tirunelveli District, Tamil Nadu, India [1]. The whole genome sequencing of the species using Ion Torrent personal genome machine (Life Technologies, Carlsbad, CA) [2] produced 458,944 raw reads having average length of 198.2 bp and
Table 2
List of virulence factors identified in the annotated genome dataset of *Bacillus cereus* 062011msu.

| SeqName | Start | Stop | Strand | Description | Length | e-Value | Sim mean |
|---------|-------|------|--------|-------------|--------|---------|----------|
| contig_1017| 240 1| − | Flagellum-specific ATP synthase | 240 | 6.88E−53 | 100 |
| contig_1017| 1466 1305| − | Flagellar assembly protein H | 162 | 3.05E−15 | 100 |
| contig_1017| 2984 1980| − | Flagellar motor switch protein F | 1005 | 0.4 | 100 |
| contig_1017| 3453 2998| − | Flagellar MS-ring protein | 456 | 1.33E−40 | 100 |
| contig_1026| 114 1106| + | UDP-glucose 4-epimerase | 993 | 0.1 | 100 |
| contig_1050| 1366 557| − | UTP-glucose-1-phosphate uridylyltransferase | 810 | 0 | 100 |
| contig_1052| 162 16 | − | Flagellar biosynthesis protein FltR | 147 | 9.6E−12 | 100 |
| contig_1052| 305 192| − | Flagellar biosynthesis protein FltQ | 114 | 4.2E−20 | 100 |
| contig_1052| 1144 503| − | Flagellar biosynthesis protein FltP | 642 | 1.91E−130 | 100 |
| contig_1052| 1792 1601| − | Flagellar motor switch protein FlmA | 192 | 7.43E−39 | 100 |
| contig_1108| 28 837| + | Flagellar motor protein MotB | 810 | 6.97E−171 | 100 |
| contig_1140| 301 1116| + | Flagellar biosynthesis regulator FltF | 816 | 0.1 | 100 |
| contig_1140| 1159 1497| + | Flagellar basal body rod protein FlgG | 339 | 6.74E−76 | 100 |
| contig_1208| 569 387| − | Flagellar hook protein FlgE | 183 | 5.15E−35 | 100 |
| contig_1208| 1752 1633| − | Flagellar basal body rod modification protein | 120 | 1.19E−10 | 100 |
| contig_1235| 13 726| + | Immune inhibitor A metalloprotease | 714 | 9.71E−152 | 100 |
| contig_1277| 720 535| − | Glycosyl transferase, group 3 family protein | 186 | 8.07E−19 | 100 |
| contig_1321| 1747 1574| − | Chemotaxis protein CheV | 237 | 1.09E−50 | 100 |
| contig_1339| 805 488| − | Nonhemolytic enterotoxin NHE | 144 | 2.77E−25 | 100 |
| contig_1358| 522 379| − | Thiol-activated cytolysin | 837 | 0.1 | 100 |
| contig_1613| 547 1092| + | Capsular polysaccharide biosynthesis protein | 264 | 4.12E−52 | 100 |
| contig_1634| 912 76| − | Hypothetical protein NEAT-type hemolysin mediated heme uptake system | 585 | 4.82E−92 | 100 |

**total size of 90,974,357 bp (90.97 MB).** The FastQC (version.0.11.5) plug-in software ([https://www.bioinformatics.babraham.ac.uk/projects/fastqc/](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)) [3] and CLC genomics workbench version 9.0.1 [4] were used for analyzing the read quality and trimming of ambiguous low quality reads. After quality assessment and trimming total 432,619 cleaned reads were obtained with an average length of 178.08 bp. The trimmed reads were assembled into 3,200 contigs with an average length of 1,611 bp and GC content of 35.3% using the denovo assembly algorithm of CLC Genomics Workbench version 9.0.1.
2.2. Genome sequence annotation and genomic data analysis

The draft genome contigs of *Bacillus cereus* 062011 msu were annotated by using the NCBI Prokaryotic Genomes Automatic Annotation Pipeline (PGAAP) [5] and Rapid Annotation of microbial genome using Subsystem Technology (RAST) version 2.0 (http://rast.nmpdr.org/) [6]. The PGAAP annotation of the isolate's genome showed total 7,061 CDS (2,301 protein coding genes and 4,760 pseudo genes) and 81 RNA genes including 66 tRNAs, 10 rRNAs and 5 ncRNAs. The annotation details were given in whole genome shotgun (WGS) project with the project accession NTMF00000000. The genomic islands are set of genes with probable horizontal origin which facilitate in the diversification, adaptation and evolution of pathogenic microbes [7,8]. The Genomic islands in our study were predicted by submitting the PGAAP generated GenBank file to the Island Viewer 4 (http://www.pathogenomics.sfu.ca/islandviewer/) [8]. Total 219 genes were clustered on 10 genetic islands.

Simultaneously the data obtained from the RAST annotation server revealed that the draft genome contains 8721 coding sequences and 472 subsystems with “Amino Acids and Derivatives” and

---

Fig. 5. Phylogenetic tree based on 16S rRNA comparison of *Bacillus cereus* 062011msu with its closely related homologs using the MEGA7 software.

Fig. 6. Circular genome map of *Bacillus cereus* 062011msu generated by DNAPlotter tool.
“Carbohydrates” were the most represented subsystem features. In addition the annotated subsystem features denoted 257 genes associated with “Virulence, Disease and Defense” including 154 genes associated with antibiotics and toxicity resistance, 52 genes associated with the synthesis of antibacterial peptides, Bacteriocins, 50 genes associated with invasion and intracellular resistance and one gene associated with adhesion. The KEGG (Kyoto Encyclopedia of Genes and Genomes) biological pathways associated with the genome of Bacillus cereus 062011msu were identified by annotating the protein coding sequences against the KEGG pathway database using the BLAST2GO program [9]. A total of 3104 sequences were mapped to 116 different KEGG pathways. Among them the pathways associated with Nucleotide metabolism, Amino acid metabolism, Metabolism of cofactors and vitamins and Carbohydrate metabolism were the most dominant KEGG pathways observed in the genome dataset. The prediction and classification of the orthologous groups associated with the Protein coding genes of Bacillus cereus 062011msu were performed by using the EggNog database (Evolutionary genealogy of genes) embedded within the BLAST2GO software [10]. The COG (Clusters of Orthologous Groups) data denoted that the cluster for “Amino acid transport and metabolism” (381 sequences) forms the largest functional group. Among the other functional groups the clusters for “Function unknown” (344 sequences), “General Function Prediction Only” (289 sequences), “Inorganic ion transport and metabolism” (281 sequences) and “Carbohydrate transport and metabolism” (261 sequences) were the highly represented categories.

Emphasizing the pathogenic nature of the strain, the virulence factors and toxic genes residing in the genome of Bacillus cereus 062011msu were further screened by annotating the coding sequences against the Virulence Factor Database (VFDB) [11] using the local BLASTX with E-value cutoff of 1E-5. A total of 1108 sequences homologous to 743 virulence factors and toxic genes were identified from the BLAST search. Among them 56 genes showed complete sequence homology (100%) with the annotated genome dataset of B. cereus 062011msu, indicating that the flagellar proteins might play regulatory role in the pathogenicity of the bacterium. The previous in vitro experiments by Mariappan et al., 2012 reported that the pathogen was sensitive to tetracycline (TET) and ciprofloxacin (CPFX) [1]. The antibiotic resistance genes present in Bacillus cereus 062011msu were screened by using the curated database, Antibiotic Resistance Genes Database (ARDB) (http://ardb.ccb.cbcb.umd.edu/) [12]. The data illustrated that the pathogen consists of total 14 crucial antibiotics resistance genes exhibiting resistance to the antibiotics like bacitracin, penicillin, fosfomycin, streptomycin, chloramphenicol, doxorubicin, fluoroquinolone, puromycin, streptomycin, beta-lactam, lincomycin and fosmidomycin, thus confirming the susceptibility of the strain to TET and CPFX.

2.3. Genome sequence comparison, 16S and 23S rRNA analysis, and genome map visualization

The closest neighboring strains for Bacillus cereus 062011msu based on the genome sequence comparison using RAST server were identified as Bacillus cereus AND1407 (score 544), Bacillus cereus MSX-D12 (score 406) and Bacillus cereus BAG3O-2 (score 387). Based on the local similarity of the aligned nucleotide sequences using the rapid sequence comparison tool BLAST [13] the genome of Bacillus cereus 062011msu exhibited 99% sequence homology with the genomes of Bacillus cereus ATCC 10987, Bacillus cereus strain M3, Bacillus cereus FRI-35, Bacillus thuringiensis serovar finitimus YBT-020, Bacillus cereus strain CC-1, Bacillus cereus NC7401, Bacillus cereus AH187 respectively. In microbial genomics research the comparison of 16S rRNA gene sequence has emerged as a reliable technique to identify new bacterial strains associated with pathogenicity and infections [14]. The deduced 16S rRNA sequence for Bacillus cereus 062011msu genome was aligned to its nearby homologs using the Clustal W multiple sequence alignment and the phylogenetic analysis was performed through the maximum likelihood method with 100 bootstrap replicates using the MEGA7 software (www.megasoftware.net/) [15]. The phylogenetic tree based on 16S rRNA sequence comparison confirmed that the pathogenic strain belongs to the species Bacillus cereus and exhibits close evolutionary relationship with B. cereus ATCC 10987 and B. cereus FRI-35 as they were clustered together as a monophyletic clade. Simultaneously we have also performed the phylogenetic analysis based on 23S rRNA gene sequence comparison using the MEGA7 software. The 16S rRNA gene derived phylogenetic tree was found to be concordant with the 23S rRNA gene tree as it also identified B. cereus ATCC 10987 as the closest evolutionary homolog of the pathogen. The complete genomic map...
of *Bacillus cereus* 062011msu representing the GC content, GC skew graphs, coordinates and coding sequence features on both forward and reverse strands obtained from RAST annotation was generated by DNAPlotter (http://www.sanger.ac.uk/science/tools/dnaplotter) [16].

**Acknowledgements**

This work was financially supported by the Bioinformatics Infrastructure Facility- Department of Biotechnology (BIF-DBT) New Delhi, India (BT/B1/04/055/2001) and Department of Science & Technology (DST-SERB), New Delhi, India (SERB/F/3766/2017–2018).

**Transparency document. Supplementary material**

Transparency document associated with this article can be found in the online version at https://doi.org/10.1016/j.dib.2017.12.054

**Appendix A. Supplementary material**

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2017.12.054

**References**

[1] S. Mariappan, S.N.M. Mydeen, S.J.R. Samuel, J. Jegathambigai, K.E. Rajan, S. Sudhakar, Bacillus sp. causing abscessation in sheep and goat population, Curr. Sci. (2012) 921–925.

[2] M.A. Quail, M. Smith, P. Coupland, T.D. Otto, S.R. Harris, T.R. Connor, A. Bertonì, H.P. Swerdlow, Y. Gu, A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers, BMC Genom. 13 (2012) 341.

[3] S. Andrews, FastQC: A Quality Control Tool for High Throughput Sequence Data, 2010.

[4] A.V. Karlyshev, V.G. Melnikov, V.A. Chistyakov, Draft genome sequence of Bacillus amyloliquefaciens B-1895, Genome Announc. 2 (2014) e00633–00614.

[5] T. Tatusova, M. DiCuccio, A. Badretdin, V. Chetvernin, E.P. Nawrocki, L. Zaslavsky, A. Lomsadze, K.D. Pruitt, M. Borodovsky, J. Ostnell, NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 44 (2016) 6614–6624.

[6] R. Overbeek, R. Olson, G.D. Pusch, G.J. Olsen, J.J. Davis, T. Disz, R.A. Edwards, S. Gerdes, B. Parrello, M. Shukla, The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST), Nucleic Acids Res. 42 (2013) D206–D214.

[7] M. Juhas, J.R. van der Meer, M. Gaillard, R.M. Harding, D.W. Hood, D.W. Crook, Genomic islands: tools of bacterial horizontal gene transfer and evolution, FEMS Microbiol. Rev. 33 (2009) 376–393.

[8] C. Bertelli, M.R. Laird, K.P. Williams, Group SFURC, B.Y. Lau, G. Hoad, G.L. Winsor, F.S. Brinkman, IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets, Nucleic Acids Res. 45 (2017) W30–W35.

[9] M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (2000) 27–30.

[10] J. Huerta-Cepas, D. Szklarczyk, K. Forslund, H. Cook, D. Heller, M.C. Walter, T. Rattei, D.R. Mende, S. Sunagawa, M. Kuhn, eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences, Nucleic Acids Res. 44 (2015) D286–D293.

[11] L. Chen, J. Yang, J. Yu, Z. Yao, L. Sun, Y. Shen, Q. Jin, VFDB: a reference database for bacterial virulence factors, Nucleic Acids Res. 33 (2005) D325–D328.

[12] B. Liu, M. Pop, ARDB—antibiotic resistance genes database, Nucleic Acids Res. 37 (2008) D443–D447.

[13] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool. J. Mol. Biol. 215 (1990) 403–410.

[14] J.E. Claridge, Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases, Clin. Microbiol. Rev. 17 (2004) 840–862.

[15] S. Kumar, G. Stecher, K. Tamura, MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets, Mol. Biol. Evol. 33 (2016) 1870–1874.

[16] T. Carver, N. Thomson, A. Bleasby, M. Berriman, J. Parkhill, DNAPlotter: circular and linear interactive genome visualization, Bioinformatics 25 (2008) 119–120.