Complete genome sequence of thermophilic Bacillus smithii type strain DSM 4216T

Bosma, Elleke Fenna; Koehorst, Jasper J.; van Hijum, Sacha A. F. T.; Renckens, Bernadet; Vriesendorp, Bastienne; van de Weijer, Antonius H. P.; Schaap, Peter J.; de Vos, Willem M.; van der Oost, John; van Kranenburg, Richard

Published in:
Standards in Genomic Sciences

Link to article, DOI:
10.1186/s40793-016-0172-8

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Complete genome sequence of thermophilic *Bacillus smithii* type strain DSM 4216<sup>T</sup>

Elleke F. Bosma<sup>1,2</sup>, Jasper J. Koehorst<sup>3</sup>, Sacha A. F. T. van Hijum<sup>4</sup>, Bernadet Renckens<sup>4</sup>, Bastienne Vriesendorp<sup>5</sup>, Antonius H. P. van de Weijer<sup>1</sup>, Peter J. Schaap<sup>3</sup>, Willem M. de Vos<sup>1</sup>, John van der Oost<sup>1</sup> and Richard van Kranenburg<sup>1,5*</sup>

**Abstract**

*Bacillus smithii* is a facultatively anaerobic, thermophilic bacterium able to use a variety of sugars that can be derived from lignocellulosic feedstocks. Being genetically accessible, it is a potential new host for biotechnological production of green chemicals from renewable resources. We determined the complete genomic sequence of the *B. smithii* type strain DSM 4216<sup>T</sup>, which consists of a 3,368,778 bp chromosome (GenBank accession number CP012024.1) and a 12,514 bp plasmid (GenBank accession number CP012025.1), together encoding 3880 genes. Genome annotation via RAST was complemented by a protein domain analysis. Some unique features of *B. smithii* central metabolism in comparison to related organisms included the lack of a standard acetate production pathway with no apparent pyruvate formate lyase, phosphotransacetylase, and acetate kinase genes, while acetate was the second fermentation product.

**Keywords:** *Bacillus smithii*, Genome sequence, Lactic acid, Thermophile, Thermophilic bacillus, Biotechnology

**Abbreviations:** DSMZ, Deutsche Sammlung von Mikroorganismen und Zellkulturen; SVs, Structural variations

**Introduction**

*Bacillus smithii* is a facultatively anaerobic, facultatively thermophilic Gram-positive bacterium, originally identified as *Bacillus coagulans* [1, 2]. Similar to its close relative *B. coagulans*, *B. smithii* has biotechnological potential, as it is able to ferment a range of carbon sources [2] into lactate and other green building block chemicals [3, 4]. The production of such green chemicals at elevated temperatures from lignocellulosic biomass has the potential to lower production costs of these chemicals. Compared to currently used mesophilic production hosts, such as Lactic Acid Bacteria or *Escherichia coli*, the amount of enzymes needed for hydrolysis of lignocellulose is ~3-fold lower around 50–60°C, which is the temperature of moderately thermophilic temperatures [5]. Furthermore, fermentation at higher temperatures decreases contamination risks and cooling costs and increases product and substrate solubility [6, 7]. In order to enable the development of *B. smithii* as a platform organism, genetic tools were recently developed for it [3, 4]. To fully exploit the biotechnological potential of this species and to gain insight into its metabolic pathways, we sequenced the genome of the *B. smithii* type strain. Reconstruction of the central metabolic pathways based on the genome reveals some remarkable differences with its close relative *B. coagulans*.

**Organism information**

**Classification and features**

*B. smithii* DSM 4216<sup>T</sup> is a motile, spore-forming, rod-shaped (0.8–1.0 by 5.0–6.0 μm [2]/0.5–1.0 by 2.0–6.0 μm, Fig. 1), facultatively anaerobic, facultatively thermophilic bacterium with wide ranges of both temperature (25–65°C) and pH (5.5–7.0) [2]. An electron micrograph of *B. smithii* DSM 4216<sup>T</sup> is shown in Fig. 1. Based on existing literature [2],
HPLC analysis [3, 4] and API-tests, it is concluded that the species is able to ferment a range of carbon sources into mainly lactate, with acetate as the major by-product and minor amounts of succinate and malate (Table 1).

In order to compare the *B. smithii* DSM 4216ᵀ genome to other fully sequenced *Bacillus* genomes, a phylogenetic tree was constructed based on 16S rRNA genes and the analysis of protein domains of *B. smithii* DSM 4216ᵀ and other currently available *Bacillus* genomes (Fig. 2) [8]. These analyses indicated that *B. smithii* is most closely related to *B. coagulans*, which is also a facultatively thermophilic species [2].

The *B. smithii* type strain DSM 4216ᵀ was isolated from cheese [1, 2], but other *B. smithii* strains have been isolated from compost [3, 9], hot spring soil [10], and a sugar beet factory [11]. It is a free-living organism that was shown to be non-cytotoxic [12]. In addition, the safety of the probiotic *B. smithii* TMBI 12 was recently reported in piglets studies [13]. Basic morphological and physiological features have been described by Nakamura et al.[2]. Genetic accessibility, a wide temperature and pH range and the ability to utilize a wide range of carbon sources in a relatively minimal medium make *B. smithii* an interesting new host for biotechnological applications [3, 4].

**Genome sequencing information**

**Genome project history**
The *B. smithii* type strain was selected based on the biotechnological relevance of the species as described above. The initial Illumina sequencing was performed in March 2012 and the genome was closed by PacBio sequencing in June 2013. The final, closed genome sequence consisting of 1 chromosome and 1 plasmid was deposited in GenBank (nr CP012024.1 and CP012025.1) and released for public access on 8 July 2015. A summary of the project information and its association with MIGS version 2.0 compliance [14] is shown in Table 2.

**Growth conditions and genomic DNA preparation**
*B. smithii* DSM 4216ᵀ was obtained from DSMZ. DNA was isolated from *B. smithii* DSM 4216ᵀ cultures grown overnight at 55 °C in 100 mL LB2 and TVMY-glucose [3] in a 250 mL Erlenmeyer. 10 mL of the cultures was harvested by centrifugation for 15 min at 4 °C and 4816 × *g*, after which DNA was isolated using the Epi-centre Master Pure Gram Positive DNA Purification kit according to the manufacturer’s protocol. DNA integrity was confirmed on a 1.0 % agarose gel and concentrations were measured using Qubit (Life Technologies), after which DNA integrity was re-evaluated by the sequencing company before sequencing.

**Genome sequencing and assembly**
The genome of *B. smithii* DSM 4216ᵀ was sequenced by BaseClear BV (NL) using Illumina HiSeq2000 mate-pair and paired-end sequencing for the initial sequencing and assembly, followed by PacBio sequencing to fully close the genome sequence. The average length of the paired-end samples was 273 bp and that of the mate-pair samples 4260 bp. The sequence reads were filtered and trimmed based on Phred quality scores, assembled into contigs using the “De Novo Assembly” option of the CLCbio Genomics Workbench version 5.0 and further
assembled into scaffolds using SSPACE Premium version 2.0 [15]. This initial sequencing resulted in 6,185,516 reads, which were assembled into 214 contigs and 27 scaffolds. The coverage of the paired-end reads was 187x and that of the mate pair reads was 311x. For gap closure, sequencing was performed using a PacBio SMRT cell and quality was again assessed based on Phred scores. PacBio sequencing resulted in 90,013 reads with an average read length of 2075 kbp and a coverage of 56x. The contigs were assembled into super-scaffolds using alignment of the PacBio reads with BLASR [16], which was then used to determine the order of and distance between the contigs using a modified SSPACE Premium version 2.3 [15]. This resulted in 5 scaffolds, after which a second PacBio run was performed, which resulted in 114,294 reads with an average length of 2775 kbp. These results were analyzed in the same way as the first PacBio-round, after which gaps in the super-scaffolds

<table>
<thead>
<tr>
<th>MIGS ID</th>
<th>Property</th>
<th>Term</th>
<th>Evidence codea</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Classification</td>
<td>Domain Bacteria</td>
<td>TAS [30]</td>
</tr>
<tr>
<td></td>
<td>Phylum</td>
<td>Firmicutes</td>
<td>TAS [31–33]</td>
</tr>
<tr>
<td></td>
<td>Class</td>
<td>Bacilli</td>
<td>TAS [34, 35]</td>
</tr>
<tr>
<td></td>
<td>Order</td>
<td>Bacillales</td>
<td>TAS [36, 37]</td>
</tr>
<tr>
<td></td>
<td>Family</td>
<td>Bacillaceae</td>
<td>TAS [37, 38]</td>
</tr>
<tr>
<td></td>
<td>Genus</td>
<td>Bacillus</td>
<td>TAS [37–39]</td>
</tr>
<tr>
<td></td>
<td>Species</td>
<td>Bacillus smithii</td>
<td>TAS [2]</td>
</tr>
<tr>
<td></td>
<td>Type strain</td>
<td>DSM 4216</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gram stain</td>
<td>Positiveb</td>
<td>TAS [2]</td>
</tr>
<tr>
<td></td>
<td>Cell shape</td>
<td>Rod</td>
<td>IDA (Fig. 1), TAS [2]</td>
</tr>
<tr>
<td></td>
<td>Motility</td>
<td>Motile</td>
<td>TAS [2]</td>
</tr>
<tr>
<td></td>
<td>Sporulation</td>
<td>Terminal or sub terminal, oval or cylindrical endospores, non-swollen to slightly swollen sporangia</td>
<td>IDA (Fig. 1), TAS [2]</td>
</tr>
<tr>
<td></td>
<td>Temperature range</td>
<td>25–65 °C</td>
<td>TAS [2]</td>
</tr>
<tr>
<td></td>
<td>Optimum temperature</td>
<td>55 °C</td>
<td>IDA</td>
</tr>
<tr>
<td></td>
<td>pH range; Optimum</td>
<td>5.5–6.8; 6.5</td>
<td>TAS [2], IDA</td>
</tr>
<tr>
<td>MIGS-6</td>
<td>Habitat</td>
<td>Type strain: cheese. Other strains: evaporated milk, canned food, compost, hot spring soil, sugar beet juice from extraction installations.</td>
<td>TAS [2, 9–11]</td>
</tr>
<tr>
<td>MIGS-6.3</td>
<td>Salinity</td>
<td>Not in 3 % NaCl (w/v)</td>
<td>TAS [2]</td>
</tr>
<tr>
<td>MIGS-22</td>
<td>Oxygen requirement</td>
<td>Facultative anaerobe</td>
<td>TAS [2]</td>
</tr>
<tr>
<td>MIGS-14</td>
<td>Pathogenicity</td>
<td>Non-pathogen</td>
<td>TAS [12, 13]</td>
</tr>
<tr>
<td>MIGS-4</td>
<td>Geographic location</td>
<td>USA</td>
<td>TAS [2, 40]</td>
</tr>
<tr>
<td>MIGS-5</td>
<td>Sample collection</td>
<td>~1946</td>
<td>TAS [2, 40]</td>
</tr>
<tr>
<td>MIGS-4.1</td>
<td>Latitude</td>
<td>Unknown</td>
<td></td>
</tr>
<tr>
<td>MIGS-4.2</td>
<td>Longitude</td>
<td>Unknown</td>
<td></td>
</tr>
<tr>
<td>MIGS-4.4</td>
<td>Altitude</td>
<td>Unknown</td>
<td></td>
</tr>
</tbody>
</table>

aEvidence codes – IDA: Inferred from Direct Assay; TAS: Traceable Author Statement (i.e., a direct report exists in the literature); NAS: Non-traceable Author Statement (i.e., not directly observed for the living, isolated sample, but based on a generally accepted property for the species, or anecdotal evidence). These evidence codes are from the Gene Ontology project.

bAs described in the species description by Nakamura et al.: “Young cells of both groups were Gram positive. With increasing age the cells became Gram variable and finally Gram negative. The KOH and aminopeptidase tests were negative, as is typical for Gram-positive organisms.”
were closed using GapFiller 1.10 [17], resulting in the final genome of 1 chromosome and 1 plasmid. Two small scaffolds (<450 bp) were found to be contaminants and removed from the data set.

Structural variations (SVs; small nucleotide polymorphism and small insertions and deletions) in the paired end and mate paired Illumina reads were compared to the PacBio scaffolds at the CMBI Nijmegen using an in-house developed tool RoVar [18]. Repeat masking of the reference sequence was done by (i) creating 30-bp fragments, (ii) aligning these fragments to the PacBio reference sequence by using BLAT [19] with a tile size of 6, and (iii) masking regions to which fragments align perfectly in multiple positions in the reference sequence. Illumina read alignment performed by BLAT with a tile size of 6 and alignment events were allowed provided that SVs were at least 4 bp from the end of a given read. SVs were used for further analysis provided that they were supported by at least 5 unique forward and 5 unique reverse reads and at most 1 % of the reads were allowed to suggest an alternative allele. A total of 14 SVs were corrected in the *B. smithii* DSM 4216\(^T\) PacBio assembly.

**Genome annotation**

The corrected PacBio assembly was subjected to RAST annotation [20] using default parameters. The following tools were used to predict gene functions (Table 4): Aragorn for tRNAs [21], RNAmmer for rRNA [22], and CRISPR-finder for CRISPR repeats and spacers [23]. The annotation was manually curated by running a BLAST of all genes and comparing starts and stops to the best hits. Via this method, also pseudogenes were manually identified.

As several pathways commonly found in bacilli were not identified by RAST in *B. smithii*, an analysis based on protein domains was performed on the *B. smithii* DSM 4216\(^T\) genome using InterProScan 5 (version

---

**Fig. 2** Phylogenetic tree based on 16S rRNA gene sequences (left) and protein domains (right). A comparison is included (horizontal lines) between the two trees, showing the position of *Bacillus smithii* DSM 4216\(^T\) relative to other *Bacillus* strains, as well as several industrially important Lactic Acid Bacterium strains. Only strains were used for which a complete genome sequence is available (as of 18 September 2014) in order to be able to perform the domain-based analysis. The 16S sequences were aligned using DECIPHER (R) [29] and the distance analysis was performed using a Jukes-Cantor correction. Phylogenetic analysis of all domains was performed by re-annotation of all proteins from selected genomes using InterProScan 5-RC7 and transformed into a absence-presence matrix. Distance was calculated using a standard Euclidean distance and clustering was performed by complete method using hclust. Tree comparison was performed by dendextend. Note that “unique” nodes between the 16S and domain-based tree are indicated with dashed lines (i.e. the order is the same but the subclustering is not). GenBank IDs of used whole genome sequences in order from top to bottom: AE016877.1, AL009126.3, CP000000.3, BA000004.3, CP012024.1, CP002472.1, CP002835.1, CP002693.1, CP001638.1, CP000557.1, CP006254.2, CP002442.1, CP002050.1, CP004008.1, CP003125.1, BA000043.1, CP000922.1, CP002222.1, CP001617.1
Table 2 Project information of the whole genome sequence of B. smithii DSM 4216T

<table>
<thead>
<tr>
<th>MIGS ID</th>
<th>Property</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIGS 31</td>
<td>Finishing quality</td>
<td>Finished</td>
</tr>
<tr>
<td>MIGS-28</td>
<td>Libraries used</td>
<td>Mate-pair (average 4,260 bp), paired-end (average 273 bp), PacBio (2,075 and 2,775 kbp)</td>
</tr>
<tr>
<td>MIGS 29</td>
<td>Sequencing platforms</td>
<td>Illumina and PacBio</td>
</tr>
<tr>
<td>MIGS 31.2</td>
<td>Fold coverage</td>
<td>Illumina paired-end: 187x, Illumina mate pair: 311x, PacBio: 56x</td>
</tr>
<tr>
<td>MIGS 30</td>
<td>Assemblers</td>
<td>CLCbio Genomics Workbench 5.0, SSPACE Premium 2.0, GapFiller 1.10</td>
</tr>
<tr>
<td>MIGS 32</td>
<td>Gene calling method</td>
<td>RAST and domain analysis</td>
</tr>
<tr>
<td>Locus Tag</td>
<td></td>
<td>BSM4216</td>
</tr>
<tr>
<td>Genbank ID</td>
<td></td>
<td>CP012024.1 (chromosome); CP012025.1 (plasmid)</td>
</tr>
<tr>
<td>GenBank Date of Release</td>
<td></td>
<td>8 July 2015</td>
</tr>
<tr>
<td>GOLD ID</td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>BIOPROJECT</td>
<td></td>
<td>PRJNA258357</td>
</tr>
<tr>
<td>MIGS 13</td>
<td>Source Material Identifier</td>
<td>Biotechnological</td>
</tr>
<tr>
<td>Project relevance</td>
<td></td>
<td>DSM 4216T</td>
</tr>
</tbody>
</table>

5RC7, 27th January 2014) (Koehorst & Van Dam, submitted for publication). This has been shown to be a powerful tool for identifying previously unknown protein functions, for example in determining microbial symbiotic interactions [8]. The domain-based annotation was compared to the manually curated RAST annotation, after which duplicates were removed and genes identified uniquely via the domain-analysis were added. In total 142 extra genes were annotated via this method, of which all except 4 were hypothetical proteins. For 209 genes, the protein domain annotation resulted in the addition of EC-numbers to the annotation that had not been assigned via RAST.

**Genome properties**

The genome of B. smithii DSM 4216T consists of a circular chromosome of 3,368,778 base pairs with a GC content of 40.8 % and a plasmid of 12,514 base pairs and a GC content of 35.9 % (Table 3). Figure 3 shows a map of the DSM 4216T plasmid and chromosome. On the chromosome, a total of 3880 genes were identified, of which 3627 were annotated as protein-coding genes, of which 81 are assigned ‘putative’ or ‘probable’ functions, 1472 are hypotheticals or genes with unknown function (38.2 %) and the remaining had a defined function. Out of the total chromosomal genes, 126 genes are pseudogenes and 94 are tRNAs, 33 are rRNA genes, 122 are genes with signal sequences for secretion and 795 are genes with a transmembrane domain (Table 4). The rRNA genes are clustered in 11 operons, which is relatively many and is thought to be linked to the capacity to grow fast in different conditions [24]. Eight of these operons were found on the forward strand and 3 on the reverse strand. Six of the operons appear to be positioned approximately opposite of each other on the two strands, while the remaining five are located very closely to the origin and to each other on the forward strand. The plasmid DNA was predicted to contain 18 genes, of which 5 have a function assigned, 11 are hypotheticals and 2 are mobile element associated proteins. The COG-distribution of genes is shown in Table 5.

**Insights from the genome sequence**

As the number of available genome sequences from thermophilic bacilli is still rather limited and B. smithii also grows at mesophilic temperatures, we compared its genome properties to those of thermophilic bacilli as well as to those of several commonly studied mesophilic bacilli (Table 6). Compared to its close relative B. coagulans, B. smithii has a slightly larger genome with a lower GC content. Compared to most mesophilic bacilli, its genome is smaller and it has a higher GC content than B. cereus but lower than B. halodurans and B. subtilis. As will be discussed in the next section, the genome content differs from its close relatives in several ways.

**Central carbon metabolism and main product pathways**

To be able to use B. smithii as a host for biotechnological purposes, it is important to understand its metabolic pathways. In the B. smithii DSM 4216T genome, all genes involved in glycolysis, gluconeogenesis, pentose phosphate pathway, TCA-cycle and glyoxylate shunt could be identified, but not the complete sets of genes for the phosphoketolase and Entner-Doudoroff pathways. Uptake systems for all sugars shown to support growth in the API-test were annotated by the RAST annotation. The organization of the xylose catabolic operon is similar to that found in B. coagulans XZLA [25]. A reconstruction of the central carbon metabolism of B. smithii DSM 4216T is shown in Fig. 4. An L-lactate dehydrogenase gene was annotated, which is in accordance with L-lactate being the major fermentation product of B. smithii [3, 4]. After RAST annotation, the methylglyoxal pathway was identified only towards D-lactate, but an in-depth analysis of protein domains also revealed the presence of all genes necessary for L-lactate
production via methylglyoxal. Based on 16S rRNA gene and complete protein domain analysis (Fig. 2), the closest relative of *B. smithii* is *B. coagulans*. However, when reconstructing the metabolic network of *B. smithii*, several remarkable differences between *B. smithii* and *B. coagulans* as well as other bacilli were observed. The most striking difference with bacilli in general is the absence of the genes coding for phosphotransacetylase and acetate kinase, which form the standard acetate production pathway in bacteria. This was confirmed by the domain-based analysis. Moreover, we also could not identify these two genes in the genome sequence of *B. smithii* strain 7_3_47FAA, which is available from a metagenome database. The fact that *B. smithii* produces significant amounts of acetate from glucose [3, 4] indicates that an alternative pathway is involved, which is...
currently being investigated. Furthermore, candidate genes for pyruvate-formate lyase, pyruvate decarboxylase and pyruvate oxidoreductase could not be found in the genomes of both DSM 4216 and 7_3_47FAA via either RAST or domain-based analysis. Therefore, \( \text{pdhc} \)-encoded pyruvate dehydrogenase complex is most likely the only enzyme responsible for the conversion of pyruvate to acetyl-CoA. This is confirmed by a \( \text{pdhA} \)-knockout strain of \( B. \) smithii strain ET 138, which is unable to grow without acetate supplementation and did not produce any acetate [4].

Another difference with \( B. \) coagulans is the lack of a catabolic \( \text{alsSD} \)-operon in \( B. \) smithii, coding for the enzymes acetolactate synthase and acetolactate decarboxylase. This is in accordance with the absence of 2,3-butanediol production [3, 4]. The anabolic acetolactate synthase small and large subunit genes \( \text{ilvBH} \) (also called \( \alpha \)-acetohydroxyacid synthase) are present. These genes are mainly involved in the isoleucine and valine biosynthetic pathways [26]. On the other hand, both an S- and an S/R-acetoin specific 2,3-butanediol dehydrogenase gene were identified in the genome. Although several alcohol dehydrogenases were found in the

**Table 4** Genome statistics of \( B. \) smithii DSM 4216\(^1 \)

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
<th>% of total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome size (bp)</td>
<td>3,381,292</td>
<td>100.0</td>
</tr>
<tr>
<td>DNA coding (bp)</td>
<td>2,799,365</td>
<td>82.8</td>
</tr>
<tr>
<td>DNA G + C (bp)</td>
<td>1,378,026</td>
<td>40.8</td>
</tr>
<tr>
<td>DNA scaffolds</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Total genes</td>
<td>3,880</td>
<td>100.0</td>
</tr>
<tr>
<td>Protein coding genes</td>
<td>3,627( ^a )</td>
<td>93.5</td>
</tr>
<tr>
<td>RNA genes</td>
<td>127</td>
<td>3.3</td>
</tr>
<tr>
<td>Pseudo genes</td>
<td>126</td>
<td>3.2</td>
</tr>
<tr>
<td>Genes in internal clusters</td>
<td>ND</td>
<td></td>
</tr>
<tr>
<td>Genes with function prediction</td>
<td>2,063</td>
<td>53.1</td>
</tr>
<tr>
<td>Genes assigned to COGs</td>
<td>3,627( ^a )</td>
<td>93.5</td>
</tr>
<tr>
<td>Genes with Pfam domains</td>
<td>2,596</td>
<td>66.8</td>
</tr>
<tr>
<td>Genes with signal peptides</td>
<td>122</td>
<td>3.1</td>
</tr>
<tr>
<td>Genes with transmembrane helices</td>
<td>795</td>
<td>20.5</td>
</tr>
<tr>
<td>CRISPR repeats</td>
<td>69</td>
<td></td>
</tr>
</tbody>
</table>

\( ^a \) This is excluding 126 pseudogenes

**Table 5** Number of genes associated with general COG functional categories

<table>
<thead>
<tr>
<th>Code</th>
<th>Value</th>
<th>% age</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>J</td>
<td>162</td>
<td>4.46</td>
<td>Translation, ribosomal structure and biogenesis</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0.00</td>
<td>RNA processing and modification</td>
</tr>
<tr>
<td>K</td>
<td>179</td>
<td>4.92</td>
<td>Transcription</td>
</tr>
<tr>
<td>L</td>
<td>160</td>
<td>4.40</td>
<td>Replication, recombination and repair</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0.03</td>
<td>Chromatin structure and dynamics</td>
</tr>
<tr>
<td>D</td>
<td>28</td>
<td>0.77</td>
<td>Cell cycle control, Cell division, chromosome partitioning</td>
</tr>
<tr>
<td>V</td>
<td>31</td>
<td>0.85</td>
<td>Defense mechanisms</td>
</tr>
<tr>
<td>T</td>
<td>125</td>
<td>3.44</td>
<td>Signal transduction mechanisms</td>
</tr>
<tr>
<td>M</td>
<td>132</td>
<td>3.63</td>
<td>Cell wall/membrane biogenesis</td>
</tr>
<tr>
<td>N</td>
<td>64</td>
<td>1.76</td>
<td>Cell motility</td>
</tr>
<tr>
<td>U</td>
<td>42</td>
<td>1.16</td>
<td>Intracellular trafficking and secretion</td>
</tr>
<tr>
<td>O</td>
<td>92</td>
<td>2.53</td>
<td>Posttranslational modification, protein turnover, chaperones</td>
</tr>
<tr>
<td>C</td>
<td>156</td>
<td>4.29</td>
<td>Energy production and conversion</td>
</tr>
<tr>
<td>G</td>
<td>174</td>
<td>4.79</td>
<td>Carbohydrate transport and metabolism</td>
</tr>
<tr>
<td>E</td>
<td>291</td>
<td>8.01</td>
<td>Amino acid transport and metabolism</td>
</tr>
<tr>
<td>F</td>
<td>74</td>
<td>2.04</td>
<td>Nucleotide transport and metabolism</td>
</tr>
<tr>
<td>H</td>
<td>107</td>
<td>2.94</td>
<td>Coenzyme transport and metabolism</td>
</tr>
<tr>
<td>I</td>
<td>94</td>
<td>2.59</td>
<td>Lipid transport and metabolism</td>
</tr>
<tr>
<td>P</td>
<td>154</td>
<td>4.24</td>
<td>Inorganic ion transport and metabolism</td>
</tr>
<tr>
<td>Q</td>
<td>70</td>
<td>1.93</td>
<td>Secondary metabolites biosynthesis, transport and catabolism</td>
</tr>
<tr>
<td>R</td>
<td>382</td>
<td>10.51</td>
<td>General function prediction only</td>
</tr>
<tr>
<td>S</td>
<td>236</td>
<td>6.49</td>
<td>Function unknown</td>
</tr>
<tr>
<td>-</td>
<td>1,321</td>
<td>36.34</td>
<td>Not in COGs</td>
</tr>
</tbody>
</table>

**Table 6** Comparison of several published complete genome sequences of the genus Bacillus

<table>
<thead>
<tr>
<th>Species/strain</th>
<th>Genome size (bp)</th>
<th>GC %(^a)</th>
<th>ORFs(^b)</th>
<th>Plasmid number</th>
<th>Growth(^c)</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>( B. ) smithii DSM 4216(^1 )</td>
<td>3,368,778</td>
<td>40.8</td>
<td>3,635</td>
<td>1</td>
<td>TT</td>
<td>This study</td>
</tr>
<tr>
<td>( B. ) coagulans DSM 1(^* )</td>
<td>3,018,045</td>
<td>47.2</td>
<td>3,437</td>
<td>0</td>
<td>TT</td>
<td>[41]</td>
</tr>
<tr>
<td>( B. ) coagulans 36D1</td>
<td>3,552,226</td>
<td>46.5</td>
<td>3,306</td>
<td>0</td>
<td>TT</td>
<td>[42]</td>
</tr>
<tr>
<td>( B. ) coagulans 2-6</td>
<td>3,073,079</td>
<td>47.3</td>
<td>2,985</td>
<td>1</td>
<td>TT</td>
<td>[43]</td>
</tr>
<tr>
<td>( A. ) flavithermus WK1</td>
<td>2,846,746</td>
<td>41.8</td>
<td>2,863</td>
<td>0</td>
<td>TT</td>
<td>[44]</td>
</tr>
<tr>
<td>( B. ) licheniformis 10-1</td>
<td>4,317,010</td>
<td>45.9</td>
<td>4,650</td>
<td>0</td>
<td>TT</td>
<td>[45, 46]</td>
</tr>
<tr>
<td>( B. ) licheniformis DSM 13(^1 )</td>
<td>4,222,748</td>
<td>46.2</td>
<td>4,286</td>
<td>0</td>
<td>TT</td>
<td>[47]</td>
</tr>
<tr>
<td>( B. ) cereus ATCC 14579</td>
<td>5,426,909</td>
<td>35.3</td>
<td>5,366</td>
<td>1</td>
<td>MP</td>
<td>[48]</td>
</tr>
<tr>
<td>( B. ) halodurans C-125</td>
<td>4,202,353</td>
<td>43.7</td>
<td>4,066</td>
<td>0</td>
<td>MP</td>
<td>[49]</td>
</tr>
<tr>
<td>( B. ) subtilis 168(^# )</td>
<td>4,214,810</td>
<td>43.5</td>
<td>4,100</td>
<td>0</td>
<td>MP</td>
<td>[50]</td>
</tr>
<tr>
<td>( G. ) thermoglucosidans TNO.09.020(^* )</td>
<td>3.75 Mb</td>
<td>43.9</td>
<td>4,300</td>
<td>0</td>
<td>TP</td>
<td>[51]</td>
</tr>
<tr>
<td>( G. ) thermodenitrificans NG80-2</td>
<td>3,550,319</td>
<td>48.9</td>
<td>3,499</td>
<td>1</td>
<td>TP</td>
<td>[52]</td>
</tr>
<tr>
<td>( G. ) kaustophilus HTA426</td>
<td>3,544,776</td>
<td>52.0</td>
<td>3,498</td>
<td>1</td>
<td>TP</td>
<td>[53, 54]</td>
</tr>
<tr>
<td>( G. ) thermoleovorans CCB.UL53.LUF5</td>
<td>3,596,620</td>
<td>52.3</td>
<td>3,887</td>
<td>0</td>
<td>TP</td>
<td>[55]</td>
</tr>
</tbody>
</table>

Currently available thermophilic \( B. \) cereus\(^* \) genomes are shown, as well as a selection of genomes of mesophilic model organisms

\( \^a \) Sequence not fully closed

\( \^b \) GC% of chromosome and plasmid combined as weighted average

\( \^c \) ORFs: Open Reading Frames as a total on chromosome and plasmid(s)

\( \# \) Open Reading Frames as a total on chromosome and plasmid(s)

\( \# \) MP: mesophile, TP: thermophile, TT: thermotolerant (grows at mesophilic as well as thermophilic temperatures)
Fig. 4 (See legend on next page.)
genome, no bifunctional acetaldehyde dehydrogenase-alcohol dehydrogenase adhE could be found, which is in accordance with the absence of alcohol production in the majority of B. smithii fermentations [3].

**Amino acid and vitamin biosynthesis pathways**

Microorganisms used for biotechnological purposes should have minimal nutrient requirements, as the addition of yeast extract, vitamins or amino acids is costly. The organisms should therefore preferably contain the pathways to synthesize vitamins, amino acids, purines and pyrimidines. In B. smithii DSM 4216T, all amino acid biosynthetic pathways could be identified. Pathways for de novo synthesis and salvage pathways of pyrimidines and purines were also identified. Complete vitamin biosynthesis pathways were identified for cobalamine, riboflavin, tetrahydrofolate, panthothenate, p-aminobenzoic acid, nicotinic acid and pyridoxal, but not for thiamine, ascorbate, pyridoxamine and D-biotin.

**Host-defense systems**

Robustness against infection is crucial for industrial microorganisms. Host-defense systems can confer such robustness, but might also hinder genetic accessibility of the organism. In the genome of B. smithii DSM 4216T, several host-defense systems are annotated: a type II-s restriction endonuclease, a 5-methylcytosine-specific restriction related enzyme, a type I restriction-modification system and a CRISPR-Cas Type I-B system. The CRISPR-Cas genes show the typical type I-B gene arrangement [27], but seem to be partly duplicated around the CRISPR locus with a second locus containing cas6, cas8a/cst1, cas7 (originally annotated as ‘CRISPR-associated negative autoregulator’) and cas5 after the CRISPR repeats. The CRISPR-finder tool [23] was used to identify CRISPR repeats and spacers in the area around the Cas-genes (bp 2,772,457-2,799,872). Three CRISPR-loci were identified (CRISPR 1, 3 and 4) as well as one questionable locus (CRISPR 2). Using CRISPRTarget [28], some of the spacers were found to have hits with potential protospacer target sequences against Bacillus sp. and B. subtilis plasmid DNA, and against Streptococcus thermophilus, Lactococcus, Enterococcus and Campylobacter phage DNA.

**Conclusions**

This report describes the complete genome sequence of Bacillus smithii type strain DSM 4216T. The species has biotechnological potential due to its efficient conversion of both C5 and C6 sugars at 55 °C to lactic acid, combined with its genetic accessibility. Its central carbon metabolism is different from its close relative B. coagulans as it lacks the alsSD operon, as well as the pta-ack acetate production pathway and the pfl gene.

**Acknowledgements**

This project was financially supported by Corbion (NL).

**Authors’ contributions**

EFB and AHPvdW performed the microbiology and molecular biology studies. BV, RK, JJK, PJS, SAFTvH and BR performed bio-informatics analyses. EFB, RK and BV were involved in the preparation and submission of the genome to NCBI. EFB drafted the manuscript, with input and revisions from JJK, PJS, SAFTvH, WdV, JvdO and RK. All authors participated in the design of the study, which was coordinated by RK, WdV and JvdO. All authors read and approved the final manuscript.

**Competing interests**

The authors declare to have no competing interests. RK and BV are employed by Corbion. EFB and AHPvdW were financially supported by Corbion.

**Author details**

1Laboratory of Microbiology, Wageningen University, Dreijenplein 10, 6703 HB Wageningen, The Netherlands. 2Present address: The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kemitorvet, Kgs. Lyngby 2800, Denmark. 3Laboratory of Systems and Synthetic Biology, Wageningen University, Dreijenplein 10, 6703 HB Wageningen, The Netherlands. 4CMBI, NCMLS, Geert-Gooteplein Zuid 26-28, 6525 GA Nijmegen, The Netherlands. 5Corbion, Akerdijk 46, 4206 CC Gorinchem, The Netherlands.

Received: 21 August 2015 Accepted: 9 August 2016 Published online: 24 August 2016