Abstract

The thermophilic ‘Geobacilli’ are important sources of thermostable enzymes and other biotechnologically relevant macromolecules. The present work reports the high quality draft genome sequences of previously unsequenced type strains of Geobacillus uzenensis (DSM 23175\textsuperscript{T}), G. thermocatenulatus (DSM 730\textsuperscript{T}) and Parageobacillus galactosidasius (DSM 18751\textsuperscript{T}). Phylogenomic analyses revealed that DSM 18751\textsuperscript{T} and DSM 23175\textsuperscript{T} represent later heterotypic synonyms of P. toebii and G. subterraneus, respectively, while DSM 730\textsuperscript{T} represents the type strain for the species G. thermocatenulatus. These genome sequences will contribute towards a deeper understanding of the ecological and biological diversity and the biotechnological exploitation of the ‘geobacilli’.

Key words: Geobacillus; Parageobacillus; Firmicutes; thermophile; phylogenomics; Illumina HiSeq sequencing

Introduction

The ‘geobacilli’ are cosmopolitan thermophilic Firmicutes that are highly adaptable and consequently have been isolated from wide range of environments, including oil wells, deserts, hot springs, compost and soils [1]. The taxonomy of these bacteria has recently been re-examined through phylogenomics, resulting in the genus Geobacillus [2] being divided into two genera: Geobacillus and Parageobacillus [3]. These genera have been the subject of increasing interest because of their ability to produce a wide range of thermostable enzymes, such as amylases, proteases, lipases, hemicellulolytic enzymes and other industrially and biotechnologically relevant macromolecules [4-5]. The increasing availability and accessibility of complete genome sequences, together with the development of tools that allow for accurate functional annotation of genomic data, are enhancing the ways in which microorganisms can be studied and characterized [6]. Furthermore, these genome sequences provide a resource for tapping into the biotechnological potential of microorganisms. Elucidating the genome sequences of type strains is especially important for resolving the taxonomic status of microorganisms [7].

Currently, the genome sequences of sixty-eight Geobacillus and sixteen Parageobacillus strains are...
publically available. These include the genome sequences of eleven and five validly described type strains of *Geobacillus* and *Parageobacillus*, respectively. The genomes of the *G. uzenensis* DSM 23175\(^T\), *G. thermocatenulatus* DSM 730\(^T\) [2] and *P. galactosidasius* DSM 18751\(^T\) [8] were paired-end sequenced using the Illumina HiSeq platform (Illumina, Inc., San Diego, CA, USA). A total of 3.6 Gb (ca. 1,060x coverage), 3.8 Gb (ca. 1,059x coverage) and 3.7 Gb (ca. 964x coverage) of reads were generated for *G. uzenensis* DSM 23175\(^T\), *G. thermocatenulatus* DSM 730\(^T\) and *P. galactosidasius* DSM 18751\(^T\), respectively. The reads were assembled using SPAdes [9], and the resulting contigs were further assembled using Multi-Draft based scaffold (MeDusa3) [10] and Mauve 2.3.1 [11]. Finally, the genomes were annotated using RAST [12] and EggNOG 4.5.1 [13] and checked for completeness using BUSCO [14]. The genome sequences were assembled to high quality draft status (between two and ten contigs) and range in size between 3.56 and 3.79 Mb, coding for between 3,783 and 4,067 proteins (Table 1). A substantially lower G+C content was observed for the *Parageobacillus* genome (41.6%) compared to the *Geobacillus* spp. (51.8 and 52.2% respectively), which represents a distinguishing feature between the two genera [3]. Assessment of the three genomes using the Firmicutes dataset indicated that all the genomes were ca. 99.4% complete. Classification of proteins into their COG functional categories based on EggNOG showed similar proportions of proteins in the different functional groups among the three strains (Figure 1), although a larger proportion of proteins involved in metabolism are present in the two *Geobacillus* isolates (Figure 1A). In particular, there are a larger proportion of proteins involved in amino acid, carbohydrate and lipid metabolism in the *Geobacillus* strains (Figure 1B), suggesting that greater metabolic versatility exists in the *Geobacillus* strains compared to *P. galactosidasius* DSM 18751\(^T\). By contrast, an elevated number of proteins (334 proteins; 8.21% of total proteins) involved in DNA replication, recombination and repair (Figure 1B) in *P. galactosidasius* DSM 18751\(^T\) compared to the other strains (246 and 244 proteins for DSM 730\(^T\) and DSM 23175\(^T\), respectively) may indicate a distinct mobilome exists in the former strain.

Maximum likelihood phylogenies were constructed on the basis of the core proteins conserved among 11 *Geobacillus* and 7 *Parageobacillus* genomes, including the 3 genomes sequenced in this study. A total of 1,355 conserved proteins were identified using Orthofinder [15], aligned using T-coffee [16], concatenated and trimmed using GBlocks [17] before the resulting alignment (296,082 amino acids in length) was used to construct a core genome maximum likelihood phylogeny using PhyML-SMS with SH-aLRT branch support method [18]. The core protein phylogeny showed that *G. thermocatenulatus* DSM 730\(^T\) clusters with three strains namely, *G. thermocatenulatus* CS-1, *G. thermocatenulatus* BCO2 and *G. thermocatenulatus* T6, in a clade previously shown to represent a distinct *Geobacillus* genomospecies [3]. *G. uzenensis* DSM 23175\(^T\) clusters with the type strain of *G. subterraneus* (DSM 13552\(^T\)). *P. galactosidasius* DSM 18751\(^T\) also clusters with the type strain of *P. toebii* (DSM 14590\(^T\)) and two other *P. toebii* strains.
Table 1. Genome features of the sequenced Geobacillus/Parageobacillus species

| Species          | Strain       | Genome size (Mb) | # Contigs | G+C (%) | # encoded proteins | # RNAs | Isolation source       | Reference |
|------------------|--------------|------------------|-----------|---------|--------------------|--------|------------------------|-----------|
| G. thermocatenulatus | DSM 730T    | 3.56             | 2         | 51.8    | 3,783              | 109    | Hot gas well (Russia)  | [2]       |
| G. uzenensis     | DSM 23175T  | 3.36             | 10        | 52.2    | 3,589              | 115    | Oil field (Kazakhstan) | [2]       |
| P. galactosidasius| DSM 18751T  | 3.79             | 6         | 41.6    | 4,067              | 127    | Compost (Italy)        | [8]       |

1 Obtained from the Bacillus Genetic Stock Centre (BGSC) at Ohio State University, USA.
2 Obtained from Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ), Leibniz, Braunschweig, Germany.

Several phylogenomic methods, including digital DNA-DNA Hybridization (dDDH) and Average Nucleotide Identity (ANI) calculations have been developed and have been shown to accurately distinguish between strains at the species level [19-20]. Pairwise BLAST-based Average Nucleotide Identity values (ANIb) were obtained using JSpecies [21], and dDDH values were calculated with the Genome-to-Genome Distance Calculator (GGDC 2.1), using formula 2 [19]. G. thermocatenulatus DSM 730T showed the highest similarity with G. thermocatenulatus T6 with an ANI value of 99.7% and dDDH of 93.6%, which far exceeds the species cut-off thresholds of 96% and 70% for ANI and dDDH, respectively. Comparison of the 16S rRNA gene sequences indicated that the gene from G. uzenensis DSM 23175T showed 99.9% sequence identity with that of G. subterraneus DSM 13552T, while the two genomes shared 99.6% ANI and 93.1% dDDH values. Furthermore, the 16S rRNA gene of P. galactosidasius DSM 18751T shared 99.3% sequence identity with that of P. toebii DSM 14590T. Phylogenomic analyses indicated that the two strains had ANI and dDDH values of 98.2% and 87.9%, respectively, both of which exceed the threshold values for species circumscription.

Based on these phylogenomic analyses, we can conclude that P. galactosidasius DSM 18751T and G. uzenensis DSM 23175T most likely represent later heterotypic synonyms of P. toebii and G. subterraneus, respectively, rather than type strains of distinct species as previously described. Conversely, we can conclusively characterize G. thermocatenulatus DSM 730T as the type strain for the species G. thermocatenulatus. Regardless of this, these genome sequences will be of additive value towards the exploration of the diversity among the geobacilli and to further explore the biotechnological potential of these Geobacillus and Parageobacillus species.

Nucleotide sequence accession numbers

The whole genome sequences have been deposited at DDBJ/EMBL/Genbank under the accession numbers NEWK00000000 (G. thermocatenulatus DSM 730T), NEWL00000000 (G. uzenensis DSM 13552T) and NDYL00000000 (P. galactosidasius DSM 18751T). The versions described in this paper are the first versions, NEWK01000000, NEWL01000000 and NDYL01000000, respectively.

Acknowledgments

WR, NK and SP were supported by the National Research Foundation (NRF) of South Africa. HA and PL were
supported by postdoctoral fellowships from the University of Pretoria, South Africa. We acknowledge support by Deutsche Forschungsgemeinschaft and Open Access Publishing Fund of Karlsruhe Institute of Technology.

Competing Interests

The authors have declared that no competing interest exists.

References

1. Zeigler DR. The *Geobacillus* paradox: why is a thermophilic bacterial genus so prevalent on a mesophilic planet? Microbiol. 2014; 160: 1-11.
2. Nazina T, Tourova TT, Poltaraus A, Novikova E, Grigoryan A, Ivanova A, et al. Taxonomic study of aerobic thermophilic bacilli: descriptions of *Geobacillus subterraneus* gen. nov., sp. nov. and *Geobacillus uzenensis* sp. nov. from petroleum reservoirs and transfer of *Bacillus* *stearothermophilus*, *Bacillus thermocatenatus*, *Bacillus thermoleovorans*, *Bacillus kaustophilus*, *Bacillus thermodenitrificans* to Geobacillus as the new combinations *G. stearothermophilus*, *G. thermocatenatus*, *G. thermoleovorans*, *G. kaustophilus*, *G. thermoleovorans* and *G. thermodenitrificans*. Int J Syst Evol Microbiol. 2001; 51: 433-446.
3. Aliyu H, Lebre PH, Blom J, Cowan DA, De Maayer P. Phylogenomic re-assessment of the thermophilic genus *Geobacillus*. Syst Appl Microbiol. 2016; 39: 527-533.
4. Hussein AH, Lisowska BK, Leak DJ. The genus *Geobacillus* and their biotechnological potential. Adv Appl Microbiol. 2015; 92: 1-48.
5. De Maayer P, Brummen PJ, Mead DA, Cowan DA. Comparative analysis of the *Geobacillus* hemicellulose utilization locus reveals a highly variable target for improved hemicellulolysis. BMC Genomics. 2014; 15: 836.
6. Loman NJ, Fallen MJ. Twenty years of bacterial genome sequencing. Nat Rev Microbiol. 2015; 13: 787-794.
7. Gao B, Gupta RS. Microbial systematics in the post-genomics era. Antonie Van Leeuwenhoek. 2012; 101: 45-54.
8. Poli A, Læzza G, Gul-Cuven R, Orlando P, Nicolaus B. *Geobacillus galactosidasi* sp. nov., a new thermophilic galactosidase-producing bacterium isolated from compost. Syst Appl Microbiol. 2011; 34: 419-423.
9. 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. J Comp Biol. 2012; 19: 455-477.
10. Bosi E, Donati B, Galardini M, Brunetti S, Sagot M-F, Lòi P, et al. MeDuSa: a multi-draft based scaffoldier. Bioinformatics. 2015; 31: 2443-2451.
11. Darling E, Mau B, Perna NT. Progressive Mauve: multiple genome alignment with gene gain, loss and rearrangement. PloS One. 2010; 5: e11147.
12. Overbeek R., Olson R, Pusch GD, Olsen GJ, Davis JJ, Díaz T, et al. The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). Nucleic Acid Res. 2014; 42: 206-214.
13. Huerta-Cepas J, Forsslund K, Szklarczyk D, Jensen JJ, von Mering C, Bork P. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol Biol Evol. 2017; [Epub ahead of print].
14. Šumakova A, Waterhouse RM, Ioannidis P, Krontsveeva EV, Zolohov EM, BUSCO; assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015; 31: 3210-3212.
15. Edmonds DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015; 16: 137.
16. Di Tommaso P, Moretti S, Xenarios I, Orobitg M, Montanyola A, Chang J, et al. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res. 2011; 39: W13-W17.
17. Castejón J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000; 17: 540-552.
18. Alford V, Longueville JF, Gascuel O. SMS: Smart Model Selection in PhyML. Mol Biol Evol. 2017; 34: 2422-2424.
19. Meier-Kolthoff JP, Aucht AF, Klenk HP, Coker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics. 2013; 14: 60.
20. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci USA. 2009; 106: 19126-19131.
21. Richter M, Rosselló-Móra R, Glickner FO, Peplies J. jSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. Bioinformatics. 2015; 32: 929-931.