Phylogenomic re-assessment of the thermophilic genus *Geobacillus*

Habibu Aliyu[1], Pedro Lebre[1], Jochen Blom[4], Don Cowan[1, 2*] don.cowan@up.ac.za, Pieter De Maayer[1, 3]

[1]Centre for Microbial Ecology and Genomics, University of Pretoria, 0028, South Africa

[2]Department of Genetics, University of Pretoria, 0028, South Africa

[3]Department of Microbiology, University of Pretoria, 0028, South Africa

[4]Bioinformatics & Systems Biology, Justus-Liebig-University Giessen, 35392 Giessen, Hesse, Germany

[*]Corresponding author at: Centre for Microbial Ecology and Genomics, University of Pretoria, 0028, South Africa.

ABSTRACT

Geobacillus is a genus of Gram-positive, aerobic, spore-forming obligate thermophiles. The descriptions and subsequent affiliations of the species in the genus have mostly been based on polyphasic taxonomy rules that include traditional sequence-based methods such as DNA-DNA hybridization and comparison of 16S rRNA gene sequences. Currently, there are fifteen validly described species within the genus. The availability of whole genome sequences has provided an opportunity to validate and/or re-assess these conventional estimates of genome relatedness. We have applied whole genome approaches to estimate the phylogenetic relatedness among the sixty-three *Geobacillus* strains for which genome sequences are currently publicly available, including the type strains of eleven validly described species. The phylogenomic metrics AAI (Average Amino acid Identity), ANI (Average Nucleotide Identity) and dDDH (digital DNA-DNA hybridization) indicated that the current genus *Geobacillus* is comprised of sixteen distinct genomospecies, including several potentially novel species. Furthermore, a phylogeny constructed on the basis of the core genes identified from the whole genome analyses indicated that the genus clusters into two monophyletic clades that clearly differ in terms of nucleotide base composition. The G+C content ranges for clade I and II were 48.8-53.1% and 42.1-44.4%, respectively. We therefore suggest that the *Geobacillus* species currently residing within clade II be considered as a new genus.

Keywords: *Geobacillus*; thermophile; Phylogenomic metrics; Average Nucleotide Identity; digital DNA-DNA hybridization

INTRODUCTION

The ultimate goal of microbial taxonomy is to unambiguously assign organisms to distinct taxa on the basis of a set of guidelines that are aimed at ascertaining the degree of relatedness among the organisms [42]. Various methods, which rely on the estimation of degree of sequence similarity between organisms, have been applied in the classification of microorganisms. Among the well-established and widely accepted methods, DNA-DNA hybridization (DDH) or re-association, together with estimates of G+C content deviation [4, 20] and 16S rRNA gene sequence analysis [4] have remained the conventional standards for microbial species circumscription and assignment of bacteria to higher taxa [26]. When a new strain shows DDH values of <70% or < 97% 16S rRNA gene sequence identity with a type strain of a given species, this new strain is judged to belong to different species [50]. The threshold for 16S rRNA gene sequence similarity for species delineation has been variously modified to a range of 98.2 and 99.0%, to accommodate discrepancies noted in the correlation of 16S rRNA sequence similarity with various DDH estimates [19, 29, 44]. Despite the discrepancies observed in species level circumscription [19, 29, 44] , the use of the 16S rRNA gene marker remains the gold standard in assignments of strains to higher taxa [52].

A number of researchers have recently made extensive and logical arguments for the re-evaluation of the guidelines that are used in microbial taxonomy [11, 49, 51]. This comes, particularly, in the light of improvements in DNA sequencing technology which have resulted in an unprecedented reduction in sequencing cost and increase in quality and quantity of sequence data [27]. In particular, Vandamme and Peeters [51] highlighted the original intention behind the application of DDH methods, which was to exploit whole genome sequences for determining the level of relatedness between microorganisms. It is argued that genomic approaches provide superior tools for species delineation and phylogeny and should be included in polyphasic taxonomic practices [51]. Two of these methods, namely average nucleotide identity (ANI) [15, 21, 40] and the genome-to-genome distance method (GGDC) [29], have been shown to be robust in estimating the ‘true’ relatedness of any set of microbial genomes. ANI values of 95–96% and *in silico* DDH values of 70% estimated using GGDC formulae have been shown to accurately estimate the DNA-DNA re-association species boundary value of 70% [15, 29, 40]. However, to date no direct genomic thresholds have been agreed upon for higher taxa classification in bacteria. Unlike laboratory-based DDH methods, phylogenomic information determined from whole genome sequences has been shown to be highly reproducible, transferable and readily available for validation by other researchers [21].

In addition to ANI and GGDC, which are derived directly from DNA sequences, several other methods, based on the conserved amino acids sequences encoded on the genomes, have been described [20, 21].

Members of the genus *Geobacillus* are Gram-positive, aerobic and spore-forming thermophiles and are frequently isolated from hot environments, including hot springs, oil wells, compost and desert soils, although they have also been isolated from more temperate environmental sources [55]. *Geobacillus* species are of biotechnological and industrial importance as they produce an array of thermostable and thermoactive biomolecules with a wide range of applications [45, 55]. The genus *Geobacillus* was proposed following the emendation of the obligately thermophilic *Bacillus* species group 5 [32]. This original genus description included six species: *G. kaustophilus*, *G. thermocatenulatus*, *G. thermodenitrificans*, *G. thermoleovorans*, *G. thermoglucosidasius* and the type species *G. stearothermophilus* [32]. Subsequently, nine additional species have been validly described or transferred to the genus: *G. caldoxylosilyticus* [13], *G. galactosidasius* [38], *G. icigianus* [5], *G. jurassicus* [34], *G. lituanicus* [22], G. *thermantarcticus* [8], *G. toebii* [46], *G. uzenensis* [32], and *G. vulcani* [33]. The genome of '*G. zalihae'* NBRC 101842 has been sequenced, but this species is not validly published [1]. Three species that were previously included in the genus: *G. debilis*, *G. pallidus* and *G. tepidamans*, have been reassigned as *Caldibacillus debilis* [8], *Aeribacillus pallidus* [31] and *Anoxybacillus tepidamans* [8], respectively. As of June 2016, a further 834 strains have been affiliated to the genus on the basis of 16S rRNA gene sequence analysis [7]. DNA sequences of the *infB*, *rpoB* and *spo0A* genes have also been shown to be useful for the affiliation of new strains in the genus *Geobacillus* [17, 23, 30]. However, the full gene sequence of *recN* has been demonstrated to be the most robust marker for assigning bacterial strains at the genus and species levels [54]. Currently, the genome sequences of sixty-three *Geobacillus* strains are available, many of which have not been classified at the species level. Here, we applied multiple phylogenomic strategies to assess the overall genomic relatedness of multiple *Geobacillus* strains and re-evaluate the current taxonomy of the genus.

MATERIALS AND METHODS

Geobacillus genomes

The genome data of sixty-three *Geobacillus* strains were retrieved from the GenBank assembly database [16] and the JGI IMG genome portal (Supplementary Table S1) [36]. These included twenty-three complete and forty draft genomes. The draft *Geobacillus* genomes were improved to high quality draft status by alignment against the closely related complete or higher quality draft genomes using the Multi-Draft based Scaffolder (MeDuSa) [3] and Mauve 2.3.1 [9]. The assembled genome sequences were subsequently structurally annotated using GLIMMER v 3.0.2 [10] as implemented in the RAST annotation pipeline [37].

Phylogenetic analyses

The GenBank files for each of the *Geobacillus* genomes from the RAST server [37] were uploaded to EDGAR 2.1 [2]. The core genome of the genus *Geobacillus* was determined in EDGAR using the BLAST Score Ratio Values (SRVs) of multiple genomes to assign orthologous gene sequences [25]. T-Coffee, which incorporates multiple aligners including Kalign [24], MAFFT [18] and MUSCLE [12], was implemented to generate high quality sequence alignments of the core genes [28]. The aligned core gene sets were concatenated using Phyutility v2.2.6 [43] and gaps were removed using the default setting in Gblocks v.0.91b [6, 48]. A maximum likelihood tree of the aligned concatenated sequences was constructed using the Mobyle server [35].The *recN* gene sequences were extracted from the sixty-three *Geobacillus* genome sequences, aligned and used to construct a maximum likelihood phylogeny as described for the core genome tree.

Phylogenomic metric calculations

Average Nucleotide Identity (ANI) and Average Amino acid Identity (AAI) values were calculated using the ani.rb and aai.rb scripts included in the enveomics package, using the two-way ANI and AAI options [41]. The reciprocal best hits results are reported here. Digital DNA-DNA hybridization (dDDH) values were calculated using the Genome-to-Genome Distance Calculator (GGDC 2.0) web server, applying formula 2 [29]. Percentage of Conserved Proteins (POCP) values, which have been used for the circumscriptions of bacterial genera [39], were estimated for the type strains of the *Geobacillus* species for which genome sequences are available, as well as the type strains of species in the closely related genera *Anoxybacillus* and *Bacillus*. POCP values were calculated as [(C1+C2)/(T1+T2)] × 100, where C1 and C2 represent the number of conserved proteins (E-value < $1e^{-5}$, alignment coverage > 50 % and amino acid identity value >40 %). POCP values of <50% were used as the threshold for delineating novel genera [39].

RESULTS

Geobacillus genomes

A total of sixty-three *Geobacillus* genomes were analysed in this study, including those of the type strains of eleven validly described *Geobacillus* species (Supplementary Table S1). The draft genomes, with the exception of *Geobacillus* sp. ZGT-1 (66 contigs), were assembled to fewer than twenty-five contigs. The genome sequences range in size from 2,630,157 (*G. stearothermophilus* ATCC 12980^T) to 3,993,793 (*G. thermoglucosidans* C56YS93) base pairs. The G+C contents of the *Geobacillus* genome sequences varied between 42.1% and 53.1% (Supplementary Table S1).

Phylogenomic analysis of the genus *Geobacillus*

A total of 1,048 orthologous genes were predicted for the sixty-three *Geobacillus* strains and *Anoxybacillus flavithermus* E13^T (used as an outgroup). The core genes were used to infer a whole genome maximum likelihood phylogeny (Figure 1). The phylogeny showed that the *Geobacillus* strains grouped into two major monophyletic clades (I and II). Clade I incorporates fifty *Geobacillus* strains, including the type strains of seven *Geobacillus* species: *G. icigianus* DSM 28325^T, *G. jurassicus* DSM 15726^T, *G. kaustophilus* DSM 7263^T, *G. stearothermophilus* ATCC 12980^T, *G. thermodenitrificans* DSM 465^T, *G. subterraneus* DSM 13552^T and *G. thermoleovorans* DSM 5366^T. '*G. zalihae*' NBRC 101842, which has not been validly described, is also included in this clade. The second clade is comprised of fourteen strains including four type strains: *G. caldoxylosilyticus* DSM 12041^T, *G. thermantarcticus* M1^T, *G. thermoglucosidans* DSM 2542^T and *G. toebii* DSM 14590^T.

A maximum likelihood phylogeny of the *recN* gene was also congruent in clustering the strains into the two major clades (Figure 2). The two clades observed in the whole-genome and *recN* phylogenies could also be distinguished on the basis of the genomic G+C contents of the strains in each clade. *Geobacillus* strains in clade I showed higher genome G+C contents (48.8 - 53.1%) than those included in clade II (42.1 - 44.4%) (Supplementary Table S1).

Phylogenomic metric analyses of the genus *Geobacillus*

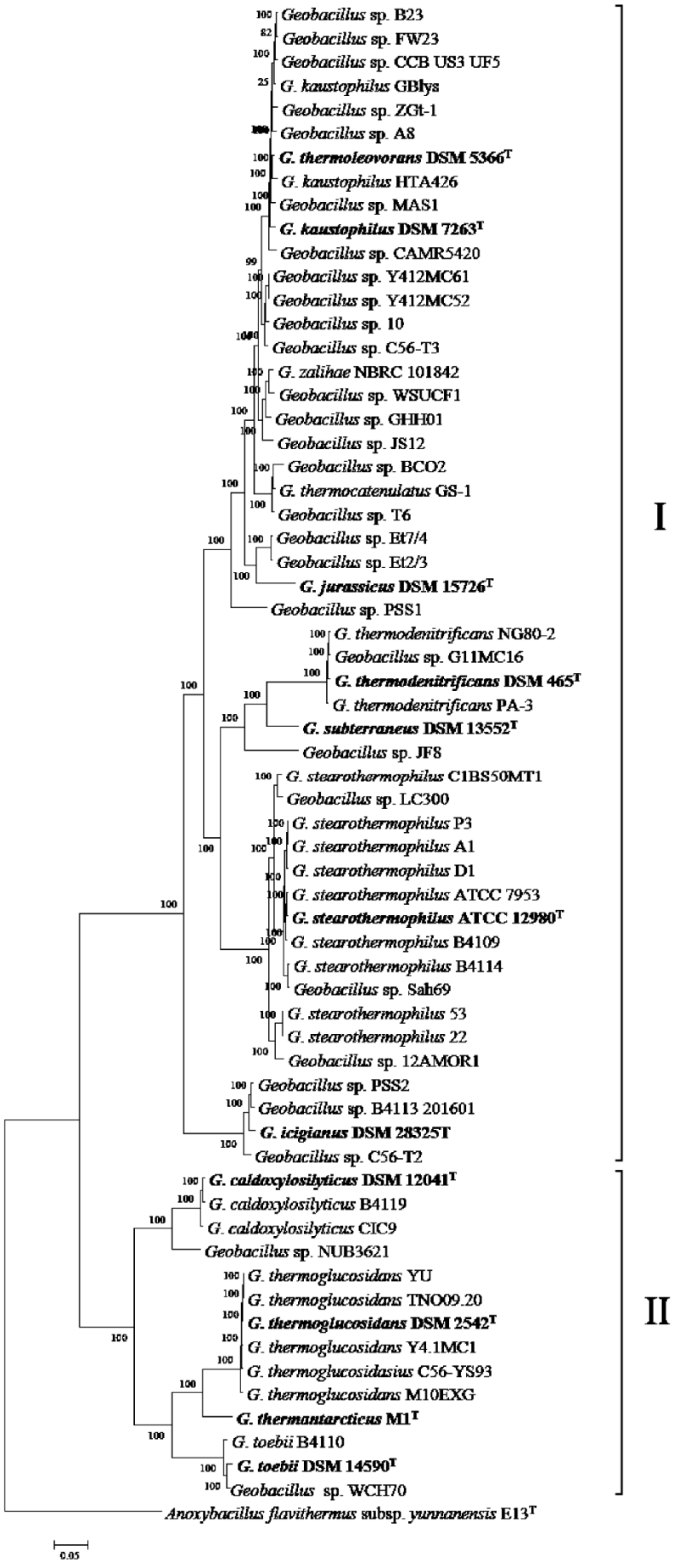

Figure 1: Whole genome phylogeny of the genus *Geobacillus*. The maximum likelihood tree was constructed based on the alignment of 1,048 concatenated core genes (total alignment length: 584,424 nucleotides) of sixty-three *Geobacillus* strains and *Anoxybacillus flavithermus* E13^T (used as outgroup). The values at the nodes indicate bootstrap values expressed as percentages of 1,000 replications while the bar length indicates 0.05 substitutions per site.

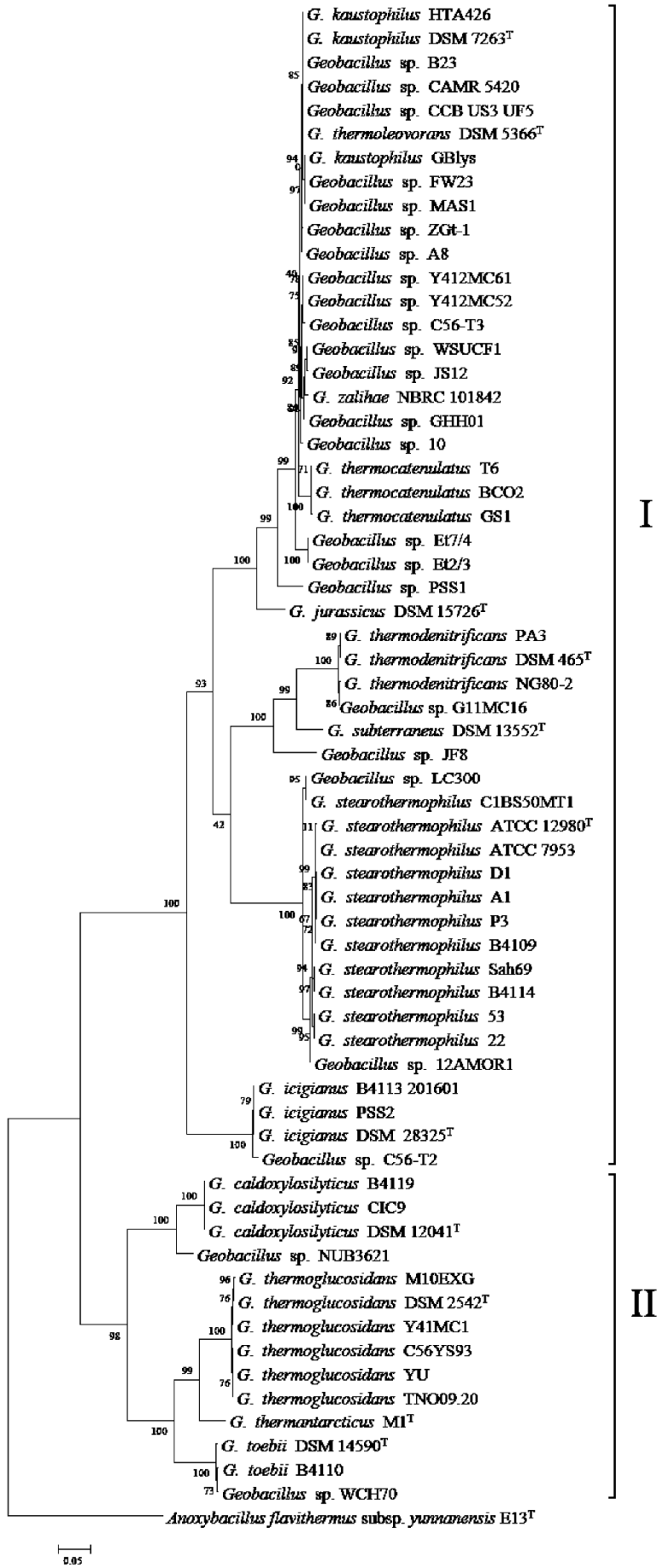
G. kaustophilus HTA426
G. kaustophilus DSM 7263T
Geobacillus sp. B23
Geobacillus sp. CAMR 5420
Geobacillus sp. CCB US3 UF5
G. thermoleovorans DSM 5366T
G. kaustophilus GBlys
Geobacillus sp. FW23
Geobacillus sp. MAS1
Geobacillus sp. ZGt-1
Geobacillus sp. A8
Geobacillus sp. Y412MC61
Geobacillus sp. Y412MC52
Geobacillus sp. C56-T3
Geobacillus sp. WSUCF1
Geobacillus sp. JS12
G. zalihae NBRC 101842
Geobacillus sp. GHH01
Geobacillus sp. 10
G. thermocatenulatus T6
G. thermocatenulatus BCO2
G. thermocatenulatus GS1
Geobacillus sp. Et7/4
Geobacillus sp. Bt2/3
Geobacillus sp. PSS1
G. jurassicus DSM 15726T
G. thermodenitrificans PA3
G. thermodenitrificans DSM 465T
G. thermodenitrificans NG80-2
Geobacillus sp. G11MC16
G. subterraneus DSM 13552T
Geobacillus sp. JF8
Geobacillus sp. LC300
G. stearothermophilus C1BS50MT1
G. stearothermophilus ATCC 12980T
G. stearothermophilus ATCC 7953
G. stearothermophilus D1
G. stearothermophilus A1
G. stearothermophilus P3
G. stearothermophilus B4109
G. stearothermophilus Sah69
G. stearothermophilus B4114
G. stearothermophilus 53
G. stearothermophilus 22
Geobacillus sp. 12AMOR1
G. icigianus B4113 201601
G. icigianus PSS2
G. icigianus DSM 28325T
Geobacillus sp. C56-T2
G. caldoxylosilyticus B4119
G. caldoxylosilyticus CIC9
G. caldoxylosilyticus DSM 12041T
Geobacillus sp. NUB3621
G. thermoglucosidans M10EXG
G. thermoglucosidans DSM 2542T
G. thermoglucosidans Y41MC1
G. thermoglucosidans C56YS93
G. thermoglucosidans YU
G. thermoglucosidans TNO09.20
G. thermantarcticus M1T
G. toebii DSM 14590T
G. toebii B4110
Geobacillus sp. WCH70
Anoxybacillus flavithermus subsp. yunnanensis E13T
I
II
0.05

Figure 2: Maximum likelihood phylogenetic tree on the basis of the *recN* gene. The phylogeny was constructed based on the *recN* gene sequences (total alignment length 1,720 nucleotides) of sixty-three *Geobacillus* strains and the outgroup strain *Anoxybacillus flavithermus* E13^{T}. The values at the nodes indicate bootstrap value expressed as percentages of 1,000 replications while the bar length indicates 0.05 substitutions per site.

Evaluation of the ANI and dDDH scores (Supplementary Table S2) revealed that *Geobacillus* clades I and II could be further partitioned into eleven and five groups (Figure 3), respectively. ANI values for members of clade I ranged from 82.4 to 100% and between 83.0 and 100% in clade II (Supplementary Table S2). The dDDH values ranged from 28.3 to 99.8% in clade I and 28.8 to 99.8% for clade II and the confidence intervals for these values do not overlap between the two clades. ANI values of 76.3 to 82.6% and dDDH values of between 23.2 and 35.9% were observed for all *Geobacillus* strains analyzed (clades I and II combined).

Group I-1 consists of sixteen strains including the type strains *G. kaustophilus* DSM 7263^T and *G. thermoleovorans* DSM 5366^T. The ANI and DDH values between DSM 7263^T and DSM 5366^T were 98.6% and 86.6%, respectively. Group I-2 includes four non-type strains and '*G. zalihae*' NBRC 101842. *G. jurassicus* DSM 15726^T is the only member of group I-6. The second largest group, group I-7, consists of thirteen strains that include the type strain of the genus, *G. stearothermophilus* ATCC 12980^T. Four strains each are clustered in groups I-8 and I-9, together with *G. icigianus* DSM 28325^{T}and *G. thermodenitrificans* DSM 465^T, respectively. *G. subterraneus* DSM 13552^T was the only strain in Group I-10. Group I-3, I-4, I-5 and I-11 did not contain any of the described type strains of validly described *Geobacillus* species (Figure 3).

Group II-1 in clade II includes *G. caldoxylosilyticus* DSM 12041^T and two non-type strains. Group II-3 consisted of six strains including the *G. thermoglucosidans* type strain, DSM 2542^T. *G. thermantarcticus* M1^T is the only strain in group II-4 while Group II-5 consisted of *G. toebii* DSM 14590^T and two non-type strains. The only group without a type strain of a validly described *Geobacillus* species in Clade II is Group II-2.

A dendrogram constructed on the basis of a distance matrix derived from the AAI values (Supplementary Table S3) showed similar clustering of the *Geobacillus* species as derived using the ANI and dDDH analyses (Supplementary Figure 1). Comparisons of the POCP values between type strains of species in the genus *Geobacillus* and selected type strains of *Anoxybacillus* and *Bacillus* species revealed that all pairwise POCP values between *Geobacillus* and *Anoxybacillus* species were greater than 50%. Considering the cut-off threshold suggested for genus circumscription using POCP, this would suggest that the compared *Geobacillus* and *Anoxybacillus* strains belong to the same genus. The highest POCP value between strains in the genus *Bacillus* and any strain in the genera *Anoxybacillus* and *Geobacillus* was 47.6%, suggesting that they do belong to two different genera. By contrast,

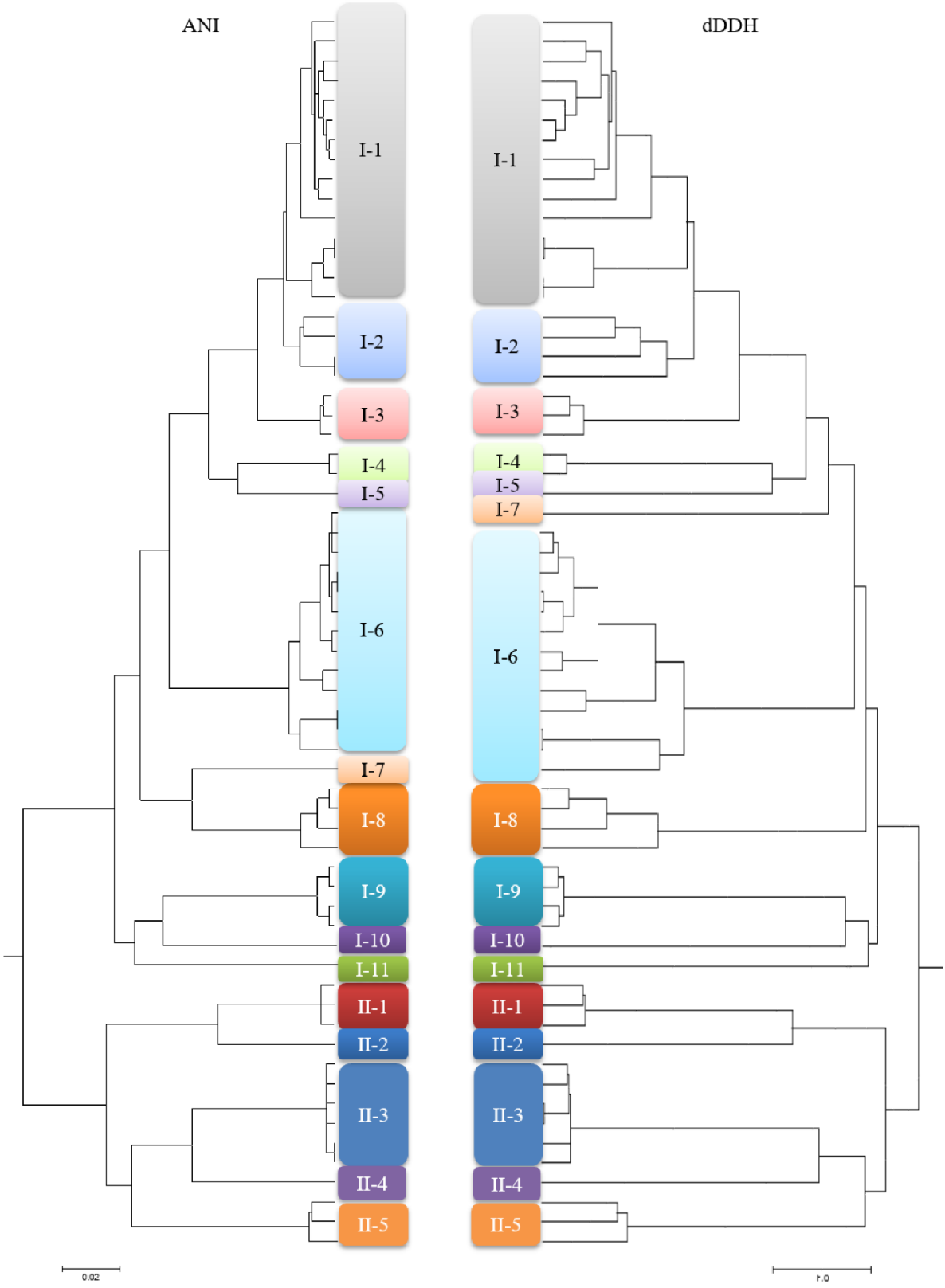

Figure 3: ANI and dDDH relationships among sixty-three strains of *Geobacillus*. The dendrogram was constructed using the distances matrices (derived from ANI and dDDH values) by using the web server DendroUPGMA [14]. The strains are numbered as follows: **Group I-1**: *G. kaustophilus* DSM 7263^{T}, HTA426 and GBlys, *Geobacillus* spp. CAMR5420,

Y4.1MC4, MAS1, FW23, A8, 10, C56-T3, Y412MC52, Y412MC61 and ZGT-1, *G. thermoleovorans* DSM 5366^{T}, B23 and CCB_US3_UF5, **Group I-2**: *Geobacillus* sp. GHH01, WSUCF1, JS12 and *G. zalihae* NBRC 101842, **Group I-3**: *G. thermocatenulatus* GS-1, *Geobacillus* sp. T6 and BCO2, **Group I-4**: *Geobacillus* sp. Et2/3 and Et7/4, **Group I-5**: *G. jurassicus* DSM 15726^{T}, **Group I-6**: *G. stearothermophilus* ATCC 12980^{T}, ATCC 7953, D1, A1, P3, B4109, B4114, C1BS50MT1, *Geobacillus* sp. LC300, 12AMOR1, 53, 22 and Sah69 **Group I-7**: *Geobacillus* sp. PSS1, **Group I-8**: *G. icigianus* DSM 28325^{T}, *Geobacillus* sp. C56-T2, PSS2 and B4113_201601, **Group I-9**: *G. thermodenitrificans* NG80-2, DSM 465^{T}, PA-3 and *Geobacillus* sp. G11MC16, **Group I-10**: *G. subterraneus* DSM 13552^{T}, **Group I-11**: *Geobacillus* sp. JF8, **Group II-1**: *G. caldoxylosilyticus* CIC9, B4119 and DSM 12041^{T}, **Group II-2**: *Geobacillus* sp. NUB3621, **Group II-3**: *G. thermoglucosidans* DSM 2542^{T} TNO-09.020, C56-YS93, M10EXG and YU and *Geobacillus* sp. Y4.1MC1, **Group II-4**: *G. thermantarcticus* M1^{T} and **Group II-5**: *G. toebii* DSM 14590^{T}, B4110 and *Geobacillus* sp. WCH70.

Figure 4: Predicted POCP values between type strains of species in the genus *Geobacillus* and type strains of selected *Anoxybacillus* and *Bacillus* species.

POCP values between *Bacillus* species ranged from 34.8 to 69.8%, suggesting they belong to different genera (Figure 4). We therefore conclude that POCP may not be reliable for circumscription of members of the family *Bacillaceae* at the genus level.

DISCUSSION

At present, the description of novel species is based on a combination of phenotypic characterisation and molecular methods, including DNA-DNA hybridization and 16S gene sequences analysis. Zeigler [53, 54] has demonstrated that *recN* gene sequence show higher resolving power in discriminating *Geobacillus* and other bacterial taxa at the genus and lower taxonomic levels compared to the 16S rRNA gene. However, using single nucleotide variant genes in the core genome, Studholme [45] has shown that the use of whole genome data results in greater resolution of the phylogenetic relationships of *Geobacillus* species than the use of single house-keeping gene phylogenies. In an effort to gain a more accurate picture of the phylogeny of the genus *Geobacillus*, we applied a range of phylogenomic approaches.

Our analysis included the genomes of eleven type strains and a further fifty-two strains which have not been classified at the species level. The phylogenomic approaches employed in this study were able to discriminate all *Geobacillus* strains at the species level. In addition, we identified four potentially novel species; *Geobacillus* sp. nov 1 (*Geobacillus* sp. Et2/3and Et7/4), *Geobacillus* sp. nov 2 (*Geobacillus* sp. JF8), *Geobacillus* sp. nov 3 (*Geobacillus* sp. PSS1) and *Geobacillus* sp. nov 4 (*Geobacillus* sp. NUB3621). '*G. zalihae*' NBRC 101842, which has not been validly described previously [1], can also be considered as a distinct species. Furthermore, an ANI value of 98.6% and dDDH value of 86.6% were observed between *G. kaustophilus* DSM7263^T and *G. thermoleovorans* DSM5366^T. These values are above the 95-96% ANI and 79% dDDH thresholds [15, 29, 40], suggesting that the two strains belong to the same species. These findings agree with Sunna et al. [47], that *G. kaustophilus* and *G. thermoleovorans* are conspecific (DDH = 84%) and are at variance with Nazina et al. [33], who reported a DDH value of 54 % for the two strains. These results further highlight the imperative of applying phylogenomic metrics in microbial taxonomy. Such metrics are robust and clearly show greater resolution than single marker genes such as 16S rRNA gene sequences.

The *recN* and core gene phylogenies, and ANI, dDDH and AAI analyses, all showed the clustering of the sixty-three *Geobacillus* strains into two distinct clades. We particularly note the considerable discrepancy in genome base composition, with a mean G+C difference of

8.34%, between the strains belonging to the two clades. These data strongly support a contention that the current genus *Geobacillus* is actually composed of two distinct genera.

CONCLUSIONS

The increasing availability of genomic information and the inherent strength of phylogenomic approaches suggest that these methods should become standard applications in species delineation and description, at least for species where weaknesses in the use of single phylogenetic marker gene are evident. Here, we have shown that phylogenomic approaches provide sufficient resolution for the accurate delineation of strains within the genus *Geobacillus* and that such methods can potentially be used to identify novel species. The distinct clustering of *Geobacillus* species into two clades, showing low genomic similarity and distinct nucleotide based compositions, suggests that the extant genus *Geobacillus* may actually consist of two distinct genera.

DESCRIPTIONS

Emended description of *Geobacillus* Nazina et al. 2001, emend. Coorevits et al. 2012

Geobacillus (Ge.o. *ba.cil´lus*. Gr. n. *Gê* the Earth; L. dim. n. *bacillus* small rod; N.L. masc. n. *Geobacillus* earth or soil small rod).

The genus comprises of the genomospecies *G. thermoleovorans*, *G. zalihae*, "*G. thermocatenulatus*", *G. jurassicus*, *G. stearothermophilus*, *G. icigianus*, *G. thermodenitrificans*, *G. subterraneus* and *Geobacillus* genomospecies 1, 2 and 3. Morphological and biochemical features as described in the emended description of *Geobacillus* by Coorevits et al. [8]. The phylogenetic positions of members of the genus is shown in Figures 1 and 2. The genomic G+C content of the "genomospecies" ranges from 48.8 to 53.1%. The type species is *Geobacillus stearothermophilus.*

Description of *G. thermoleovorans comb. nov.*

The description of *G. thermoleovorans* comb. *nov.* is identical to that of the genus and to the description proposed by Sunna et al. [47]. The species includes strains from both *G. thermoleovorans* and *G. kaustophilus.*

Description of *Geobacillus* genomospecies 1

As delineation of the strains within group I-4 of clade I could be determined using AAI, ANI and dDDH (Figures 3 and Supplementary Figure S1), it is proposed to designate a novel genomospecies 1, represented by strains Et2/3 and Et7/4.

Description of *Geobacillus* genomospecies 2

As delineation of the strains within group I-7 of clade I could be determined using AAI, ANI and dDDH (Figures 3 and Supplementary Figure S1), it is proposed to designate a novel genomospecies 2, represented by strain PSS1.

Description of *Geobacillus* genomospecies 3

As delineation of the strains within group I-11 of clade I could be determined using AAI, ANI and dDDH (Figures 3 and Supplementary Figure S1), it is proposed to designate a novel genomospecies 3, represented by strain JF8.

Description of *Parageobacillus gen. nov.*

Parageobacillus (Pa.ra.ge.o.ba.cil'lus. Gr. prep. Para, beside or alongside of; n. Gê the Earth; L. dim. n. bacillus small rod; M.L. masc. n. Parageobacillus, a genus nearest to *Geobacillus).*

As delineation of strains within clade II (Figures 1 and 2) is possible based on several genome-based metrics highlighted in this work, it is proposed to designate a novel genus *Parageobacillus*. The genus incorporates five "genomospecies", *P. caldoxylosilyticus, P. thermoglucosidans, P. thermantarcticus, P. toebii* and *Parageobacillus* genomospecies 1 (NUB3621). The phylogenetic positions of members of the genus is shown in Figures 1 and 2. The genomic G+C contents of the "genomospecies" ranges from 42.1 to 44.4%. The type species is *Parageobacillus thermoglucosidans.*

Description of *Parageobacillus caldoxylosilyticus comb. nov.*

Basonym: *Saccharococcus caldoxylosilyticus* Ahmad et al. 2000; *Geobacillus caldoxylosilyticus* Fortina et al. 2001.

The description of *Parageobacillus caldoxylosilyticus comb. nov*. is identical to that given for the new genus and to the description given by Fortina et al. [13].

Description of *Parageobacillus thermoglucosidans comb. nov.*

Basonym: *Bacillus thermoglucosidasius* Suzuki et al. 1983; *Geobacillus thermoglucosidasius* Nazina et al. 2001; *Geobacillus thermoglucosidans* Coorevits et al. 2012.

The description of *Parageobacillus thermoglucosidans comb. nov*. is identical to that given for the new genus and to the emended description given by Coorevits et al. [8].

Description of *Parageobacillus thermantarcticus comb. nov.*

Basonym: *Bacillus thermantarcticus* Nicolaus et al. 2002 (*Bacillus thermoantarcticus* [sic] Nicolaus et al. 1996); *Geobacillus thermantarcticus* Coorevits et al. 2012.

The description of *Parageobacillus thermantarcticus comb. nov*. is identical to that given for the new genus and to the description given by Coorevits et al. [8].

Description of *Parageobacillus toebii comb. nov.*

Basonym: *Geobacillus toebii* Sung et al. 2002.

The description of *Parageobacillus toebii comb. nov*. is identical to that given for the new genus and to the emended description given by Coorevits et al. [8].

Description of *Parageobacillus* genomospecies 1

As delineation of the strains within group II-2 of clade II could be determined using AAI, ANI and dDDH (Figures 3 and Supplementary Figure S1), it is proposed to designate a novel genomospecies 3, represented by strain NUB3621.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the University of Pretoria (Habibu Aliyu – University of Pretoria Postdoctoral Fellowship funding), National Research Foundation (Pieter De Maayer – Research Career Advancement Fellowship, Grant # 91447) and the University of Pretoria Genomic Research Institute for funding this study.

REFERENCES

1. Abd Rahman RN, Leow TC, Salleh AB, Basri M. 2007. Geobacillus zalihae sp. nov., a thermophilic lipolytic bacterium isolated from palm oil mill effluent in Malaysia. BMC Microbiol. 7: 77

2. Blom J, Kreis J, Spänig S, Juhre T, Bertelli C, et al. 2016. EDGAR 2.0: an enhanced software platform for comparative gene content analyses. Nucleic Acids Res.: gkw255

3. Bosi E, Donati B, Galardini M, Brunetti S, Sagot M-F, et al. 2015. MeDuSa: a multi-draft based scaffolder. Bioinformatics 31: 2443-51

4. Brenner DJ, Staley JT, Krieg NR. 2005. Classification of procaryotic organisms and the concept of bacterial speciation. In Bergey's manual® of systematic bacteriology, pp. 27-32: Springer

5. Bryanskaya AV, Rozanov AS, Slynko NM, Shekhovtsov SV, Peltek SE. 2015. Geobacillus icigianus sp. nov., a thermophilic bacterium isolated from a hot spring. Int. J. Syst. Evol. Microbiol. 65: 864-9

6. Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17: 540-52

7. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, et al. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 42: D633-42

8. Coorevits A, Dinsdale AE, Halket G, Lebbe L, De Vos P, et al. 2012. Taxonomic revision of the genus Geobacillus: emendation of Geobacillus, G. stearothermophilus, G. jurassicus, G. toebii, G. thermodenitrificans and G. thermoglucosidans (nom. corrig., formerly 'thermoglucosidasius'); transfer of Bacillus thermantarcticus to the genus as G. thermantarcticus comb. nov.; proposal of Caldibacillus debilis gen. nov., comb. nov.; transfer of G. tepidamans to Anoxybacillus as A. tepidamans comb. nov.; and proposal of Anoxybacillus caldiproteolyticus sp. nov. Int. J. Syst. Evol. Microbiol. 62: 1470-85

9. Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5: e11147

10. Delcher AL, Bratke KA, Powers EC, Salzberg SL. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23: 673-79

11. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32: 1792-97

12. Figueras MJ, Beaz-Hidalgo R, Hossain MJ, Liles MR. 2014. Taxonomic affiliation of new genomes should be verified using average nucleotide identity and multilocus phylogenetic analysis. Genome announc 2: e00927-14

13. Fortina MG, Mora D, Schumann P, Parini C, Manachini PL, Stackebrandt E. 2001. Reclassification of Saccharococcus caldoxylosilyticus as Geobacillus caldoxylosilyticus (Ahmad et al. 2000) comb. nov. Int. J. Syst. Evol. Microbiol. 51: 2063-71

14. Garcia-Vallvé S, Palau J, Romeu A. 1999. Horizontal gene transfer in glycosyl hydrolases inferred from codon usage in Escherichia coli and Bacillus subtilis. Mol. Biol. Evol. 16: 1125-34

15. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int. J. Syst. Evol. Microbiol. 57: 81-91

16. Jenuth J. 2000. The NCBI. Publicly available tools and resources on the Web. Methods Mol Biol 132: 301 - 12

17. Kapralou S, Fabbretti A, Garulli C, Gualerzi CO, Pon CL, Spurio R. 2009. Characterization of Bacillus stearothermophilus infA and of its product IF1. Gene 428: 31-35

18. Katoh K, Standley DM. 2014. MAFFT: iterative refinement and additional methods. In Multiple Sequence Alignment Methods, pp. 131-46: Springer

19. Kim M, Oh H-S, Park S-C, Chun J. 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int. J. Syst. Evol. Microbiol. 64: 346-51

20. Konstantinidis K, Tiedje J. 2005. Towards a genome-based taxonomy for prokaryotes. J. Bacteriol. 187: 6258 - 64

21. Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. U.S.A. 102: 2567-72

22. Kuisiene N, Raugalas J, Chitavichius D. 2004. Geobacillus lituanicus sp. nov. Int. J. Syst. Evol. Microbiol. 54: 1991-5

23. Kuisiene N, Raugalas J, Chitavichius D. 2009. Phylogenetic, inter, and intraspecific sequence analysis of spo0A gene of the genus Geobacillus. Curr. Microbiol. 58: 547-53

24. Lassmann T, Sonnhammer EL. 2005. Kalign–an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics 6: 298

25. Lerat E, Daubin V, Moran NA. 2003. From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. PLoS Biol. 1: E19

26. Logan NA, Berge O, Bishop AH, Busse HJ, De Vos P, et al. 2009. Proposed minimal standards for describing new taxa of aerobic, endospore-forming bacteria. Int. J. Syst. Evol. Microbiol. 59: 2114-21

27. MacLean D, Jones JD, Studholme DJ. 2009. Application of next-generation sequencing technologies to microbial genetics. Nat. Rev. Microbiol. 7: 287-96

28. Magis C, Taly J-F, Bussotti G, Chang J-M, Di Tommaso P, et al. 2014. T-coffee: tree-based consistency objective function for alignment evaluation. In Multiple Sequence Alignment Methods, pp. 117-29: Springer

29. Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M. 2013. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics 14: 1-14

30. Meintanis C, Chalkou KI, Kormas KA, Lymperopoulou DS, Katsifas EA, et al. 2008. Application of rpoB sequence similarity analysis, REP-PCR and BOX-PCR for the differentiation of species within the genus Geobacillus. Lett. Appl. Microbiol. 46: 395-401

31. Minana-Galbis D, Pinzon DL, Loren JG, Manresa A, Oliart-Ros RM. 2010. Reclassification of Geobacillus pallidus (Scholz et al. 1988) Banat et al. 2004 as Aeribacillus pallidus gen. nov., comb. nov. Int. J. Syst. Evol. Microbiol. 60: 1600-4

32. Nazina T, Tourova T, Poltaraus A, Novikova E, Grigoryan A, et al. 2001. 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 thermocatenulatus, Bacillus thermoleovorans, Bacillus kaustophilus, Bacillus thermodenitrificans to Geobacillus as the new combinations G. stearothermophilus, G. th. Int. J. Syst. Evol. Microbiol. 51: 433-46

33. Nazina TN, Lebedeva EV, Poltaraus AB, Tourova TP, Grigoryan AA, et al. 2004. Geobacillus gargensis sp. nov., a novel thermophile from a hot spring, and the reclassification of Bacillus vulcani as Geobacillus vulcani comb. nov. Int. J. Syst. Evol. Microbiol. 54: 2019-24

34. Nazina TN, Sokolova D, Grigoryan AA, Shestakova NM, Mikhailova EM, et al. 2005. Geobacillus jurassicus sp. nov., a new thermophilic bacterium isolated from a high-temperature petroleum reservoir, and the validation of the Geobacillus species. Syst. Appl. Microbiol. 28: 43-53

35. Néron B, Ménager H, Maufrais C, Joly N, Maupetit J, et al. 2009. Mobyle: a new full web bioinformatics framework. Bioinformatics 25: 3005-11

36. Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, et al. 2014. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Res. 42: D26-D31

37. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, et al. 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 42: D206-D14

38. Poli A, Laezza G, Gul-Guven R, Orlando P, Nicolaus B. 2011. Geobacillus galactosidasius sp. nov., a new thermophilic galactosidase-producing bacterium isolated from compost. Syst. Appl. Microbiol. 34: 419-23

39. Qin Q-L, Xie B-B, Zhang X-Y, Chen X-L, Zhou B-C, et al. 2014. A proposed genus boundary for the prokaryotes based on genomic insights. J. Bacteriol. 196: 2210-15

40. Richter M, Rosselló-Móra R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. U.S.A. 106: 19126-31

41. Rodriguez-R LM, Konstantinidis KT. 2016. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. Rep. 2167-9843, PeerJ Preprints

42. Rossello-Mora R. 2012. Towards a taxonomy of Bacteria and Archaea based on interactive and cumulative data repositories. Environ. Microbiol. 14: 318-34

43. Smith SA, Dunn CW. 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics 24: 715-16

44. Stackebrandt E, Ebers J. 2006. Taxonomic parameters revisited: tarnished gold standards. In Microbiol. Today, pp. 152–55

45. Studholme DJ. 2015. Some (bacilli) like it hot: genomics of Geobacillus species. Microbial biotechnology 8: 40-48

46. Sung MH, Kim H, Bae JW, Rhee SK, Jeon CO, et al. 2002. Geobacillus toebii sp. nov., a novel thermophilic bacterium isolated from hay compost. Int. J. Syst. Evol. Microbiol. 52: 2251-5

47. Sunna A, Tokajian S, Burghardt J, Rainey F, Antranikian G, Hashwa F. 1997. Identification of Bacillus kaustophilus, Bacillus thermocatenulatus and Bacillus strain HSR as members of Bacillus thermoleovorans. Syst. Appl. Microbiol. 20: 232-37

48. Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56: 564-77

49. Thompson CC, Amaral GR, Campeao M, Edwards RA, Polz MF, et al. 2015. Microbial taxonomy in the post-genomic era: rebuilding from scratch? Arch. Microbiol. 197: 359-70

50. Tindall BJ, Rossello-Mora R, Busse HJ, Ludwig W, Kampfer P. 2010. Notes on the characterization of prokaryote strains for taxonomic purposes. Int. J. Syst. Evol. Microbiol. 60: 249-66

51. Vandamme P, Peeters C. 2014. Time to revisit polyphasic taxonomy. Antonie Van Leeuwenhoek 106: 57-65

52. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, et al. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat. Rev. Microbiol. 12: 635-45

53. Zeigler DR. 2003. Gene sequences useful for predicting relatedness of whole genomes in bacteria. Int. J. Syst. Evol. Microbiol. 53: 1893-900

54. Zeigler DR. 2005. Application of a recN sequence similarity analysis to the identification of species within the bacterial genus Geobacillus. Int. J. Syst. Evol. Microbiol. 55: 1171-79

55. Zeigler DR. 2014. The Geobacillus paradox: why is a thermophilic bacterial genus so prevalent on a mesophilic planet? Microbiology 160: 1-11