Antonie van Leeuwenhoek
https://doi.org/10.1007/s10482-020-01494-3

ORIGINAL PAPER

The assessment of leading traits in the taxonomy of the *Bacillus cereus* group

Mariano A. Torres Manno · Guillermo D. Repizo · Christian Magni · Christopher A. Dunlap · Martín Espariz

Received: 6 October 2020 / Accepted: 23 October 2020

Abstract *Bacillus cereus* sensu lato strains (*B. cereus* group) are widely distributed in nature and have received interest for decades due to their importance in insect pest management, food production and their positive and negative repercussions in human health. Consideration of practical uses such as virulence, physiology, morphology, or ill-defined features have been applied to describe and classify species of the group. However, current comparative studies have exposed inconsistencies between evolutionary relatedness and biological significance among genomospecies of the *B. cereus* group. Here, the combined analyses of core-based phylogeny and all versus all Average Nucleotide Identity values based on 2116 strains were conducted to update the genomospecies circumscriptions within *B. cereus* group. These analyses suggested the existence of 57 genomospecies, 37 of which are novel, thus indicating that the taxonomic identities of more than 39% of the analyzed strains should be revised or updated. In addition, we found that whole-genome in silico analyses were suitable to differentiate genomospecies such as *B. anthracis, B. cereus* and *B. thuringiensis*. The prevalence of toxin and virulence factors coding

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10482-020-01494-3) contains supplementary material, which is available to authorized users.

M. A. Torres Manno · C. Magni · M. Espariz (✉)
Laboratorio de Biotecnología e Inocuidad de los Alimentos, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Municipalidad de Granadero Baigorria, Sede Suipacha 590, Rosario, Santa Fe, Argentina
e-mail: espariz@ibr-conicet.gov.ar

M. A. Torres Manno · C. Magni · M. Espariz
Laboratorio de Fisiología y Genética de Bacterias Lácticas, Instituto de Biología Molecular y Celular de Rosario (IBR - CONICET), sede FCByF - UNR, Rosario, Santa Fe, Argentina

M. A. Torres Manno · M. Espariz
Área Estadística y Procesamiento de Datos, Departamento de Matemática y Estadística, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Rosario, Argentina

G. D. Repizo
Departamento de Microbiología, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario (UNR), Rosario, Argentina

G. D. Repizo
Laboratorio de Resistencia bacteriana a antimicrobianos, Instituto de Biología Molecular y Celular de Rosario (IBR), sede FCByF - UNR, Rosario, Santa Fe, Argentina

C. A. Dunlap
United States Department of Agriculture, Crop Bioprotection Research Unit, National Center for Agricultural Utilization Research, Agricultural Research Service, 1815 North University Street, Peoria, IL 61604, USA

Published online: 12 November 2020

genes in each of the genomospecies of the *B. cereus* group was also examined, using phylogeny-aware methods at wide-genome scale. Remarkably, Cry and emetic toxins, commonly assumed to be associated with *B. thuringiensis* and emetic *B. paranthracis,* respectively, did not show a positive correlation with those genomospecies. On the other hand, anthrax-like toxin and capsule-biosynthesis coding genes were positively correlated with *B. anthracis* genomospecies, despite not being present in all strains, and with presumably non-pathogenic genomospecies. Hence, despite these features have been so far considered relevant for industrial or medical classification of related species of the *B. cereus* group, they were inappropriate for their circumscription. In this study, genomospecies of the group were accurately affiliated and representative strains defined, generating a rational framework that will allow comparative analysis in epidemiological or ecological studies. Based on this classification the role of specific markers such as Type VII secretion system, cytolysin, bacillolysin, and siderophores such as petrobactin were pointed out for further analysis.

Keywords Average nucleotide identity · Multilocus sequence analysis · *Bacillus thuringiensis* · *Bacillus anthracis* · Comparative genomics · Phylogeny-aware linear regression models

Introduction

Bacillus cereus sensu lato (here referred as *B. cereus* group) is part of the widely distributed family *Bacillaceae* (phylum *Firmicutes*). This family is composed of at least 78 genera of mostly spore-forming Gram-positive, facultative anaerobic, chemoorganotrophic, rod-shaped bacteria. The *B. cereus* group encompasses a diverse array of pathogenic strains adapted to a broad range of hosts (Liu et al. 2017b). *B. anthracis,* the etiological agent of anthrax, and *B. cereus,* a causative agent of emetic and diarrheal human food-poisoning are the most infamous (Raymond and Bonsall 2013; Méric et al. 2018). On the other hand, *B. thuringiensis*, a pathogen of invertebrate organisms that primarily infects nematodes, is important for insect pest management (Zheng et al. 2017). Furthermore, thermotolerant strains of the *B. cereus* group, which were sporadically associated with food poisoning, have been recently circumscribed to the *B. cytotoxicus* species (Guinebretière et al. 2013). Nevertheless, strains of *B. cereus* group are also known to be adapted to diverse environmental niches such as soil or plants and even have been used as probiotic in mammals (Hong et al. 2005; Stenfors Arnesen et al. 2008). In this context, plasmid mobilization in this group of bacteria has contributed to improve species fitness, and thereby its adaptability to environmental niches (Patiño-Navarrete and Sanchis 2017).

Almost three decades ago, all members of the *Bacillaceae* family were considered as part of the *Bacillus* genus, with *B. subtilis* and *B. cereus* constituting the main clade of bacilli (Mandic-mulec et al. 2015). Continuous taxonomic revisions are still shaping the family, although some groups remain challenging even for specialists. *B. cereus* group species were defined based on horizontally acquired-genes such as *cry*, and those encoding for the anthrax toxin or capsule synthesis (Maughan and Van der Auwera 2011; Méric et al. 2018), or other physiological or morphological considerations (Guinebretière et al. 2008). As expected, the systematic species classification based on phenotypic similarities has resulted in incoherencies between evolutionary relatedness and biological significance (Maughan and Van der Auwera 2011). In agreement, phylogenetic analyses based on conserved signature indels, gene markers or whole genome sequences have suggested that *B. cereus* group is composed of non-monophyletic species with a common evolutionary origin (Alcaraz et al. 2010; Bhandari et al. 2013). Moreover, it was postulated that the entire group should be moved to a new genus, even though historical and practical constrains limit its implementation (Bhandari et al. 2013).

The *B. cereus* group was historically circumscribed within three clades that show a high degree of sexual isolation resulting in distinctive phylogenetic patterns that enable better clustering among them (Didelot et al. 2009; Patiño-Navarrete and Sanchis 2017). Clade 1 comprises all the *B. anthracis* and several *B. cereus* and *B. thuringiensis* strains, whereas Clade 2 consists of *B. cereus* and the majority of *B. thuringiensis* strains. Clade 3, which shows the greatest phylogenetic diversity, contains *B. mycoides*, *B. weihenstephanensis* and other species (Didelot et al. 2009; Zwick et al. 2012). Guinebretière and coworkers

have defined seven major phylogenetic groups (I–VII) among *B. cereus* group strains using both genetic and phenotypic criteria (Guinebretière et al. 2008). This distribution seems to be correlated with adaptation to particular thermal niches that could have restricted ecological opportunities for gene horizontal transfer (Méric et al. 2018), thereby enabling speciation (Shapiro et al. 2016). Two groups were preferentially associated with cold thermal niches (II and VI). Group II is mainly composed of *B. wiedmannii, B. thuringiensis* and *B. cereus*, whereas Group VI consists of *B. mycoides* and *B. weihenstephanensis* strains (Guinebretière et al. 2008; Miller et al. 2016). Both groups are frequent residents of "environmental" sources (Guinebretière et al. 2008). Group II was initially associated with food poisoning (Carlin et al. 2010; Guinebretière et al. 2010) but nowadays it is known that may have important industrial applications (Lazarte et al. 2018; Wu et al. 2019). Group VI strains are frequently isolated from food sources (Beno et al. 2019). Additionally, some Group VI strains are capable of synthesizing emetic toxin (cereulide) to levels that could lead to emetic intoxication during temperature abuse scenarios (Guérin et al. 2017). On the other hand, a moderate thermotolerant phenotype is associated with strains of Group VII which is composed of *B. cytotoxicus* (Carlin et al. 2010; Ceuppens et al. 2013). Mesophilic *B. pseudomycoides* strains belong to Group I, separated from the other rhizoidal colony-forming *B. mycoides* strains (Guinebretière et al. 2008). *B. cereus* and *B. thuringiensis* strains are spread over intermediate groups II, III, IV and V (Carlin et al. 2010; Ceuppens et al. 2013). Group IV strains are frequently isolated from a wide range of foods around the world, produce enterotoxins, and also have been isolated in conjunction with diarrheal illness cases (Guinebretière et al. 2008; Guinebretière et al. 2010; Amor et al. 2018; Johler et al. 2018; Miller et al. 2018). Conversely, *B. cereus* clinical isolates are located in Group III where almost all emetic *B. cereus* and *B. anthracis* strains are clustered (Carlin et al. 2010; EFSA BIOHAZ Panel 2016). Finally, strains belonging to Group II and V were recently proposed to be *B. wiedmannii* and *B. toyonensis* species, respectively (Jiménez et al. 2013; Miller et al. 2016).

In recent years, the advance of sequencing technologies and powerful whole-genome comparative tools aided bacterial species classification (Rosselló-Móra 2012; Whitman 2014). Average nucleotide identity (ANI) calculation has been one of the most widely used and accepted tools to determine species boundaries as well as to confirm isolate identifications (Chun et al. 2018; Ciufo et al. 2018). Remarkably, some tools like FastANI have emerged, allowing accurate and fast estimations of pairwise ANI values among bacterial genomes (Jain et al. 2018). Whole-genome multilocus sequence analysis (MLSA) is also a reproducible, reliable, and highly informative approach to infer phylogenetic relationships among prokaryotes (Rong and Huang 2014; Chun et al. 2018). Using these modern methods, Liu et al. (2015) analyzed 224 *B. cereus* group strains and determined that they could be classified into 30 species, with 19–20 of them representing putative novel ones. Notably, the authors concluded that *B. thuringiensis* and *B. cereus* should be merged into a single species. More recently, Bazinet (2017) constructed a well-supported phylogeny of 498 *B. cereus* group strains based on whole-genome data. Consistencies with the three-clade and seven-group classification systems were found. However, six strains did not fit into one of these groups and the inclusion of two new ones corresponding to "*B. manliponensis*" and a division of Group III was recommended (Bazinet 2017). Importantly, eleven new species ("*B. gaemokensis*", "*B. bingmayongensis*", *B. paranthracis, B. pacificus, B. tropicus, B. albus, B. mobilis, B. luti, B. proteolyticus, B. nitratireducens,* and *B. paramycoides*) have been recently proposed by using genome-wide systematic analysis (Jung et al. 2010; Liu et al. 2014; Liu et al. 2017a). On the contrary, in a more recent and broad study, Carroll et al. have proposed defining a new species named "*Bacillus mosaicus*". This new species would group those recently defined species *B. albus, B. mobilis, B. pacificus, B. paranthracis,* and *B. tropicus* but also the pathogenic *B. anthracis* as well as *B. wiedmannii* species defined by themselves in a previous study (Miller et al. 2016; Carroll et al. 2020). To define this species the authors used an genomo-species ANI threshold of 92.5% (Carroll et al. 2020) that was substantially lower than the 95% they used to define *B. wiedmannii* (Miller et al. 2016) or the suggested by Jain et al. (2018). In this disrupting approach, the authors proposed that only five species describe the diversity of 96.69% of the *Bacillus cereus* group strains. On the other hand, this proposed framework makes use of subspecies names and biovar

epithets to account for the heterogeneity of clinically and industrially important phenotypes (Carroll et al. 2020).

Accuracy in species assignations extremely impacts the way industrial strains are selected, grown, approved for commercialization, and finally commercialized, due to the fact that such assignations, implicitly or not, are used to predict bacterium safety and performance (Gevers et al. 2005). Moreover, correct species assignation showed to be a prerequisite for truthful comparative and predictive analysis (Torres Manno et al. 2019). Hence, the European Food Safety Authority (EFSA) Panel on Biological Hazards (2016) has recommended the application of whole-genome sequencing to provide unambiguous identification of strains used as biopesticides as well as to assist safety assessment by characterizing *B. cereus* group outbreaks. To generate an adequate framework to classify and define *B. cereus* group genomospecies, we performed in silico genome-wide functional and phylogenetic analyses of 2,116 strains with available genome sequences. Additionally, phylogeny-aware methods based on linear regression models were applied to examine the correlation of the presence of genes encoding toxins and virulence factors with each genomospecies of the *B. cereus* group. Our study aims to reconcile molecular, ecological, and evolutionary data to develop an operational, useful, and predictive classification system for genomospecies of the *B. cereus* group. Also, a statistical correlation between genomospecies and genes associated with relevant phenotypes, as well as the identification of unique genes conserved in the genomospecies with significant genome coverage were performed.

Materials and methods

Genome sequences acquisition and strains selection

Genomic sequence data sets as well as predicted coding sequences of *B. cereus* group strains and *Bacillus* spp. were retrieved from GenBank (ftp.ncbi.nlm.nih.gov/genomes/) using Download Genomes tool (https://github.com/torresmanno/Download_Genomes). Datasets were composed of sequences submitted until April 3rd of 2018. Reference strains were defined based on RefSeq or EZBioCloud databases (Yoon et al. 2017). 16S ribosomal gene alignment of reference *B. cereus* group strains were performed using BLASTN (Blast + 2.7.1) searches (Johnson et al. 2008) against a local database constructed with the downloaded genomes. Those strains for which their 16S rRNA genes showed a blast result with identity percentages $\leq$ 98.6% or query coverage $\leq$ 90% were discarded. Assembly accession and proposed species assignations of selected genomes are detailed in Supplementary Table S1.

Phylogenetic tree construction

The Maximum likelihood phylogenetic trees were constructed as reported in Espariz et al. (2016) with minor modifications. Briefly, the genes present in all analyzed strains (common ancestral genes) were identified by BLASTN (Blast + 2.7.1) searches (Johnson et al. 2008) using an E-value of 10^{-30}. Common ancestral genes were individually aligned by Clustal Omega V1.2.2 (Sievers et al. 2011) and trimmed using GBlock 0.91b (Talavera and Castresana 2007). To contra-select potential paralogues, coverage and identity percentage cut-offs were set at 70% using GeM-Pro tool (Torres Manno et al. 2019). Then, aligned genes that exceeded these thresholds were concatenated using Python3 package AMAS 0.98 (Borowiec 2016). Locus tags and descriptions of common ancestral genes of reference strains used in this study are listed in Supplementary Table S2. In order to remove poor informative sequence regions, those highly similar were not included in the analyses. Finally, phylogenetic relationships of strains were inferred using RAxML 8.2.12 software (Stamatakis 2014) and the GTR substitution model with Gamma distribution. The inferred tree reliabilities were evaluated by bootstrapping with 100 replicates (Stamatakis 2014). The resulting dendrograms were displayed and annotated using iTOL (Letunic and Bork 2011).

ANI estimation

ANI values of *B. cereus* group strains were calculated with FastANI V1.1 (Jain et al. 2018) program using default parameters. ANI threshold value for species circumscription was set at 96% as previously suggested (Lee et al. 2016; Liu et al. 2017a). Genomospecies were defined when 99.9% of the ANI values of

the strains that compose a monophyletic group were ≥ 96%.

Representative strain selection

In order to define representative strains, distances among common ancestral genes of each Clade were used as correlated variables in a PCA analysis. Then, medoid strains, those strains that have their linearly uncorrelated variables closer to the centroids of the PCA, were defined as representatives. In case that more than one medoid strain was found, the assembly level was used as quality criteria in order to select the best one.

Genome annotations and protein-coding gene enrichment statistical analysis

All *B. cereus* group genomes were de novo annotated with Prokka (Seemann 2014) version 1.12-beta (arguments: -kingdom Bacteria -genus Bacillus). In order to determine associations between each protein-coding gene and the different genomospecies, enrichment statistical analyses were conducted with PhyloLM V2.6 (Levy et al. 2018). The multiple comparisons for the statistical tests were corrected with Benjamini–Hochberg approach (Benjamini and Hochberg 1995).

Analysis of COG category content

Presence/absence matrixes of Prokka-predicted protein-coding gene were generated using Roary version 3.12.0 (parameters: minimum percentage identity threshold = 95%, MCL inflation = 1.5, and Float = 90%) (Page et al. 2015). Those predicted proteins encoded in 90% of the strains under analysis were taken as the soft-core. The COG category of each soft-core protein was determined with the NCBI Batch web CD-search v3.17 Tool (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi, Marchler-Bauer et al. 2017). Finally, COG enrichment statistical analyses were conducted as described above using "Poisson_GEE" method.

Analysis of presence/absence of toxins and virulence factors

Toxin and virulence protein sequences of AtxA, Lef, CapA, CapB, CapC, CapD, CapE, PagA, HasA, Cya, BceT, CerA, CerB, CesA, CesB, CesC, CesD, Clo, CytK1, CytK2, Cry, EntA, EntFM, HblA, HblB, HblC, HblD, HlyII, HlyR, InhA1, InhA2, NheA, NheB, NheC, PlcA, PlcB, and PlcR were used as query in TBLASTN or BLASTP (Blast + 2.7.1) searches using a coverage ≥ 70% and identity ≥ 50%. Accession number of query sequences are available in Supplementary Table S1 except for *B. thuringiensis* delta-endotoxin that are listed in http://www.lifesci.sussex.ac.uk/home/Neil_Crickmore/Bt/. Genome sequences or Prokka-predicted protein sequences of strains under study were used to construct BLAST databases. Finally, trait enrichment statistical analyses were conducted as described above using "logistic_IG10" method.

Gene gain/loss analysis of selected species

Genomospecies in Clades 1 and 2 with greater than 60 strains were selected for the determination of the genes gained/loss at each phylogenetic node. All available genomes were used for all selected genomospecies, except for *B. cereus* and *B. thuringiensis*. In these cases, the MLST scheme described by Rooney et al. was used to limit the number of genomes up to 2 members per MLST sequence type (Rooney et al. 2009). The 773 strains used in the analysis are listed in Supplementary Table S1. The gain/loss analysis was generated with the genome comparator function implemented under BIGSdb version 1.16.3 (Jolley and Maiden 2010). A complete genome served as a reference strain for each genomospecies and was used to BLAST all genomes in the set. The determination of genes gained or lost by the different clades was based on a 90/10 comparison. The change in genes was determined at each phylogenetic node in the tree, the change is reported as the number of genes found in > 90% of the genomes on that side of the node, and < 10% of the genomes from the opposite side of the node. Finally, those traits that were found enriched were analyzed statistically over all available genomes with PhyloLM V2.6 (Levy et al. 2018) as described above using "logistic_IG10" method.

Results

Updating the genomospecies boundaries within the *B. cereus* group

In order to broaden the knowledge regarding the physiology and genetics of *B. cereus* group genomospecies, a consensus framework that establishes their boundaries should be concretized. With this aim, a whole-genome analysis of the 2049 strains annotated at Genbank database as *B. cereus* group species was performed. In addition, 67 *Bacillus* spp. whose 16S ribosomal encoding genes share more than 98.7% identity with any of the 21 type strains of *B. cereus* group listed in Table 1 were included in the analysis. First, ten core genes ubiquitous to all strains were selected in order to determine the phylogenetic relationships of the 2116 strains. The resulting phylogenetic tree supported the existence of Clades 1 and 2, already described for the *B. cereus* group (Didelot et al. 2009; Zwick et al. 2012; Liu et al. 2015; Bazinet 2017), whereas members of Clade 3 displayed a non-monophyletic composition (Fig. S1). In fact, the latter group of strains shared an average ANI value of 89.6 ± 6.1% that contrasts with the 94.7 ± 1.9% and 97.0 ± 1.1% ANI values shared among Clade 1 or 2 members, respectively. These facts are in concordance with the previously reported higher diversity of Clade 3 (Didelot et al. 2009; Zwick et al. 2012) and challenge the actual meaning of this clade. In addition, six clusters were not fully resolved using the aforementioned ten gene sequences (Fig. S1). It has been recently proposed that a minimum number of 31 genes should be used in the classification of genera or higher taxa by means of phylogenomic treeing (Chun et al. 2018). Hence, with the aim of better defining genomospecies circumscriptions inside the *B. cereus* group, more confident phylogenetic analyses were performed.

Clade 2 is composed of strains of B. thuringiensis and the opportunistic pathogen B. cereus sensu stricto

As was previously reported, *B. cereus* group strains submitted to NCBI as *B. thuringiensis* or *B. cereus* were dispersed in all *B. cereus* clades (Liu et al. 2015; Bazinet 2017). For that reason, the rational, unambiguous, and operational definition of boundaries for both genomospecies represents a huge taxonomic challenge. Besides their wide dispersion among clades, both type strains *B. cereus* ATCC 14579^{T} and *B. thuringiensis* ATCC 10792 ^{T} are located in Clade 2 (Fig. S1). These strains are supposed to be phenotypically distinguished only by the presence of intracellular protein crystals (Rasko et al. 2005). However, DNA–DNA hybridization assays as well as genetic studies including multilocus sequence typing (MLST), fluorescent amplified fragment length polymorphism analysis (AFLP) and rep-PCR fingerprinting have shown that strains of these species are not distinguishable (Rasko et al. 2005; Han et al. 2006; Patiño-Navarrete and Sanchis 2017). Moreover, foundational genomic studies of Rasko et al. (2005), which were supported by recent reports, have suggested that *B. thuringiensis* and *B. cereus* are actually a single genomospecies (Liu et al. 2015; Bazinet 2017). In order to improve the taxonomic description of Clade 2 genomospecies, the phylogenetic history of the 910 members of the Clade was reconstructed by a genome-wide MLSA. 676 and 3736 protein encoding-genes were found to be present in 100% and 99% of the strains, constituting the core and extended core, respectively. Common ancestral genes were defined as core genes that are also encoded in outgroup strains (Espariz et al. 2016). Then, *B. anthracis* Ames and *B. mycoides* ATCC 6462^{T} strains were used as outgroups to select 74 common ancestral genes (Supplementary Table S2). The concatenated alignment of these genes was used to perform an MLSA that showed that *B. cereus* ATCC 14579^{T} and *B. thuringiensis* ATCC 10792 ^{T} were located in well separated branches (Fig. 1a). The newly genomovars proposed by Baek et al. (2019) on the basis of wide-genome scale approaches, *B. thuringiensis* gv. *thuringiensis* and *B. thuringiensis* gv. *cytolyticus*, were also identified in our analysis. Recently, Liu et al. (2017a) have used a 96% ANI value threshold in order to define 9 novel species of the *B. cereus* group. Furthermore, the physiological and biochemical characteristics of the strains support these new assignations. Hence, a 96% ANI threshold was used to analyze the relationships of strains inside *B. cereus* ATCC 14579^{T} and *B. thuringiensis* ATCC 10792 ^{T} branches. Results indicated that 99% of the strains of each branch satisfied that species criterion (Table 1). However, 66% of Clade 2 strains share an ANI value higher than 96% with the type strain of its neighbor genomospecies. In fact, *B. cereus* ATCC 14579^{T} and *B. thuringiensis*

Table 1 ANI value analysis of proposed genomospecies

Type/Genomospecies	N	Representatives	Quantile	
			1%	0.10%
B. wiedmannii FSL W8-0169	124	NMSW16	96.1	96.0
B. mobilis 0711P9-1	19	B4088	96.0	96.0
B. albus N35-10-2	6	MOD1_Bc206	97.8	97.8
Genomospecies #1	17	RIVM_BC485	98.9	98.9
Genomospecies #2	6	AFS013362	98.0	98.0
Genomospecies #3	2	F2404B_79	99.1	99.1
Genomospecies #6	40	AFS098222	96.3	96.1
Genomospecies #8	5	AFS095574	99.4	99.4
B. anthracis Ames Ancestor	177	K3	97.1	96.7
B. paranthracis Mn5	101	BDRD_ST26	96.4	96.3
B. tropicus N24	61	FT9	96.1	95.7
B. pacificus EB422	38	M3	96.7	96.6
Genomospecies #9	12	AFS025007	96.9	96.9
Genomospecies #10	4	BGSC_4BL1	99.0	99.0
Genomospecies #13	2	100374	97.5	97.5
Genomospecies #14	2	B4082	99.6	99.6
Genomospecies #15	19	AFS095575	97.3	97.3
B. cereus ATCC 14579	569	C1L	96.4	96.1
B. thuringiensis ATCC 10792	341	AFS075683	96.0	95.7
B. toyonensis BCT-7112	228	VD115	96.5	96.3
B. pseudomycoides DSM 12442	67	AFS040159	98.0	97.9
B. gaemokensis JCM 15801	2	JCM_15801	100.0	100.0
Genomospecies #19	8	AFS083043	96.2	96.1
Genomospecies #20	2	AFS023182	99.9	99.9
Genomospecies #22	43	AFS099912	97.3	97.0
Genomospecies #23	5	AFS014408	99.7	99.7
B. mycoides ATCC 6462; *B. weihenstephanensis* NBRC 101238	54	VD021	96.4	96.1
B. nitratireducens 4049	68	BAG1X1_3	97.0	96.9
B. paramycoides NH24A2	8	AFS053592	96.9	96.9
B. proteolyticus TD42	3	AFS036423	98.4	98.4
Genomospecies #25	2	INRA_SL	99.3	99.3
Genomospecies #28	4	GOE10	99.8	99.8
Genomospecies #30	14	VD118	96.8	96.8
Genomospecies #31	6	AFS079368	96.7	96.7
Genomospecies #32	3	AFS081508	97.3	97.3
B. cytotoxicus NVH 391-98	14	NVH_391_98	98.8	98.8
B. luti TD41	3	AFS058404	96.6	96.6
Genomospecies #33	5	AFS030140	96.6	96.6
Genomospecies #34	6	AFS087218	97.1	97.1
Genomospecies #35	2	AFS039342	99.9	99.9
Genomospecies #36	2	AFS070861	99.4	99.4
Genomospecies #37	7	AFS015896	99.3	99.3

N: Number of strains; * Representatives were defined based on a PCA analysis. Singletons "*B. bingmayongensis*" FJAT-1383^{T} and *B. manliponensis* JCM 15802^{T} are not shown

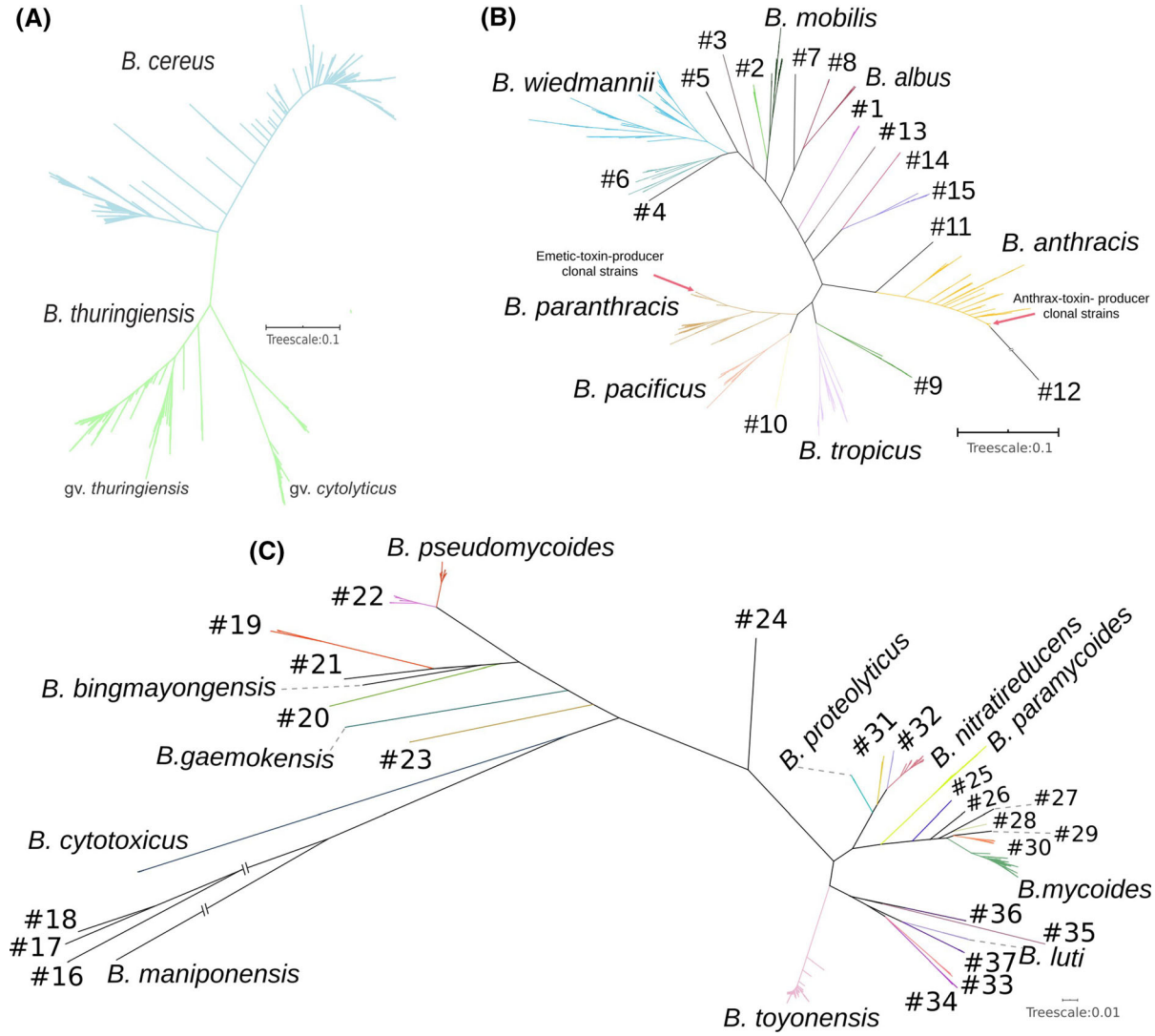

Fig. 1 Phylogenomic tree of *B. cereus* group strains. 74, 78 or 80 common ancestral genes were identified in all studied strains by BLASTN searches using an E-value of 10^{-30} and *B. cereus* ATCC 14579^{T} (**a**), *B. anthracis* Ames (**b**), or *B. mycoides* ATCC 6462^{T} (**c**) genes as query, respectively. Genes were individually aligned and trimmed and those that share $\geq$ 70% query coverage and $\geq$ 70% identity percentage were selected. The resulting final alignments comprising 51,253 (**a**), 51,150 (**b**), and 90,638 (**c**) residues were used to infer the phylogenetic relationships of *B. cereus* sensu lato group strains. Trees were constructed with RAxML algorithm (Stamatakis 2014) and displayed and annotated using iTOL (Letunic and Bork 2011). The inferred tree reliabilities were evaluated by bootstrapping with 100 replicates (Stamatakis 2014). Locus tags and descriptions of *B. cereus* ATCC 14579^{T}, *B. anthracis* Ames, and *B. mycoides* ATCC 6462^{T} genes used to infer phylogeny are listed in Supplementary Table S2

ATCC 10792 ^{T} type strains share ANI 96.7% between them. To avoid overlapping between genomospecies regarding their ANI values, the threshold should be increased up to 97.09% which, in sum with the above observations, reflect the high genomic relationship among the strains of Clade 2. On the other hand, if the threshold was set at 95.51%, members of Clade 2 would be considered a single genomospecies. Therefore, as was highlighted previously, tuning ANI thresholds for genomospecies circumscription is a daunting task when involving highly phylogenetically related strains (Ciufo et al. 2018), such as those of

Clade 2. Nevertheless, in this study, *B. cereus* sensu stricto and *B. thuringiensis* genomospecies were considered as different clusters to enable comparative assets regarding their leading traits.

Clade 1 hosts B. anthracis and other vertebrate pathogen strains

Clade 1 contains all *B. anthracis* and *B. wiedmannii* strains but also strains submitted to NCBI as *B. thuringiensis* and *B. cereus*. As aforementioned, phylogenomic analyses have shown that *B. thuringiensis,* as well as *B. cereus* sensu stricto, are members of Clade 2 and therefore those located in Clade 1 belong to other genomospecies. Hence, in order to better describe the taxonomy of the 640 strains of the Clade 1, an MLSA was performed. For this, 78 common ancestral genes were selected from the 3518 and 1368 extended core and core genes of Clade 1, respectively (Supplementary Table S2). Based on the resulting tree, 15 new genomospecies should be defined in order to better represent Clade 1 species circumscriptions (Fig. 1b). Moreover, 19 and 6 strains clustered together with type strains of recently defined species *B. mobilis* and *B. albus*, respectively (Liu et al. 2017a). As was previously reported, *B. anthracis* strains associated with anthrax disease have clonal structure (Didelot et al. 2009; Patiño-Navarrete and Sanchis 2017). In the phylogenetic tree depicted in Fig. 1b, the 113 clonal *B. anthracis* strains clustered together with the reference strain *B. anthracis* Ames, six more divergent *B. anthracis* isolates, and putative misnamed *B. thuringiensis* and *B. cereus* strains defining the *B. anthracis* genomospecies. The presence of phylolgenetically-related *B. thuringiensis* and *B. cereus* strains that share phenotypic characteristics with canonical *B. anthracis* strains was previously reported (Maughan and Van der Auwera 2011). Furthermore, emetic *B. cereus* strains cluster together within the *B. paranthracis* clade (Fig. 1b) agreeing with their suggested clonal structure (Didelot et al. 2009; Patiño-Navarrete and Sanchis 2017). Finally, it is noteworthy that ten strain clusters and five singletons not bearing type strains, and composed of putative misnamed *B. thuringiensis, B. cereus,* and/or *B. anthracis* strains were also observed in Clade 1 (Fig. 1b). In order to define representative strains, distances among common ancestral genes were used as correlated variables in a PCA analysis. Then, medoid strains, those strains that have their linearly uncorrelated variables closer to the centroids of the PCA, were defined as representatives (Fig. S2). In Table 1 are listed representatives of type-lacking genomospecies as well as type-containing genomospecies. Finally, in order to validate the species circumscriptions based on the MLSA, all vs all ANI values were computed for Clade 1 strains. As expected, 99.9% of the strains of each putative genomospecies branch satisfied the 96% species criterion (Table 1).

Unbounding the Clade 3

Members of *B. pseudomycoides*, *B. toyonensis, B. mycoides*, and *B. cytotoxicus* were described in the literature as Clade 3-species (Bazinet 2017), whereas type or reference strains of the recently defined "*B. bingmayongensis", "B. gaemokensis", "B. manliponensis", B. proteolyticus, B. nitratireducens, B. luti* and *B. paramycoides* species were found dispersed among them in a phylogenetic tree constructed using 21 common ancestral genes (Fig. S1). This fact is a consequence of the arbitrary assignation of Clade 3 members and suggests that this group of bacteria should be further divided into monophyletic groups (genomospecies) or indeed clades. We then decided to analyze the phylogenetic relationship of these 566 strains in order to better define the genomospecies contained in it. As expected, the estimated number of extended core (862) and core (156) genes found were markedly lower than that for strains within each Clade 1 or 2. Then, 80 common ancestral genes were selected using as outgroup *B. anthracis* Ames and *B. cereus* ATCC 14579^{T} strains (Supplementary Table S2). The phylogenic tree constructed using these genes suggested the existence of at least 22 new genomospecies, 8 out of them singletons (Fig. 1C). In addition, all vs all ANI value computations support the genomospecies circumscriptions suggested by our MLSA analysis (Table 1).

In summary, as a result of our classification, 279 strains should be considered as member of *B. anthracis, B. cereus, B. mycoides, B. pseudomycoides, B. thuringiensis, B. toyonensis,* or *B. wiedmannii* genomospecies. In addition, 218 strains showed to be clustered within 24 type-lacking genomospecies and 307 putative misnamed strains to the recently defined species *B. albus, B. luti, B. mobilis, B. nitratireducens, B. pacificus, B. paramycoides, B. paranthracis, B. proteolyticus,* or

B. tropicus. Finally, 15 single-strain species were identified including the two type strains "*B. bingmayongensis*" FJAT-1383 and *B. manliponensis* JCM 15802.

Main features associated with controversial species of the *B. cereus* group

Comparative genomic studies that evaluate the correlation of gene contents or phenotypes among strains have the mandatory requirement that those microorganisms should be accurately labeled at the species level. However, some important *B. cereus* group species, as *B. thuringiensis* or *B. anthracis* have been historically defined by its ability to produce insecticidal crystal proteins (Stenfors Arnesen et al. 2008) or produce anthrax toxin (Kolstø et al. 2009), respectively. As those phenotypes are associated with plasmid-encoded genes, the correlation of their presence or absence have been recently evaluated among genomospecies rather than species, as more informative procedure (EFSA BIOHAZ Panel 2016; Bazinet 2017; Zheng et al. 2017; Baek et al. 2019). A genomospecies could be defined as a monophyletic group that can be differentiated from other using genomic methods such as ANI. Taking advantage of the update in *B. cereus* group genomospecies circumscription hereby provided, we decided to use a phylogeny-correcting approach in order to evaluate whether the occurrence of *cry* and virulence factor encoding genes (detailed in Fig. 2) correlate with a particular genomospecies or clade of the group. We observed, as expected, that *cry* was significantly present in Clade 2 isolates ($p = 2.5\ 10^{-3}$) and absent in Clade 1 strains ($p = 7.0\ 10^{-3}$), the latter mostly including vertebrate pathogens (Patiño-Navarrete and Sanchis 2017; Zheng et al. 2017). However, similar percentages of *B. cereus* sensu stricto or *B. thuringiensis* strains encode Cry toxins (Fig. 2), and therefore no significant correlation with *B. thuringiensis* genomospecies was observed (Table S3). This clearly indicates that phenotypic analyses are very inappropriate procedures for species identification as was previously suggested (Liu et al. 2015; Bazinet 2017; Patiño-Navarrete and Sanchis 2017). Regarding the pathogenic characteristics of strains contained in the whole Clade 2, we found that the cytotoxin K2 gene (present in the 94% of the Clade 2 strains, Fig. 2) was the solely virulence factor showing a positive correlation

Fig. 2 Toxins and virulence factors in *B. cereus* group strains. ▶ The presence of main toxins and virulence proteins was determined by TBLASTN and BLASTP searches using a coverage ≥ 70% and identity ≥ 50%. *B. anthracis*-associated proteins (Anthrax related): AtxA, Lef, CapABCDE, PagA, HasA, Cya; Diarrheal toxin: BceT; Cereolysin: CerAB; Emetic toxin: CesABCD; Cereolysin O; Cytotoxin K1: CytK1; Cytotoxin K2: CytK2; Enterotoxin: EntAFM; Hemolysin BL: HblABCD; Haemlolysin II: HlyII,HlyR; Immune inhibitor A: InhA1,InhA2; Nonhamolytic enterotoxin: NheABC; Phospholipase C: PlcABR

($p = 2.5\ 10^{-16}$), agreeing with that reported by Carroll (2017). Nevertheless, among Clade 2 strains, cytotoxin K2 and diarrheal toxin virulence factors showed to be enriched in *B. cereus* sensu stricto with respect to *B. thuringiensis* genomospecies ($p < 0.05$). Clade 1 strains were observed to positively correlate with *B. anthracis*-associated genes, but also cereolysin O, and haemolysin II ($p < 0.05$, Table S3). When compared among Clade 1 strains, it was observed that *B. anthracis* genomospecies was positively associated with *B. anthracis*-associated genes ($p < 0.05$, Table S3), as expected. On the other hand, *B. paranthracis* genomospecies, even comprising the emetic clonal strains, did not correlate with emetic toxins, even though it was present in 24% of the isolates (Fig. 2). Furthermore, the latter genomospecies negatively correlated with haemolysin II and haemolysin BL ($p < 10^{-3}$, Table S3). Conversely, it was observed that *B. tropicus* and *B. wiedmannii* positively correlated with cytotoxin K2; and diarrheal toxins and haemolysin BL genes, respectively ($p < 0.05$, Table S3). Regarding strains that did not cluster in Clade 1 or 2, we found that the probiotic genomospecies *B. toyonensis*, mainly isolated from environmental or plant samples, was positively correlated with the presence of diarrheal toxins as well as phospholipase C ($p < 0.05$, Table S3). *B. mycoides* isolates, that were associated with food sources, were observed to positively correlate with the presence of cereolysin O and phospholipase C encoding genes ($p < 10^{-2}$, Table S3). As observed for other oversampled groups, while cytotoxin K1 coding genes were identified in all *B. cytotoxicus* genomes (Fig. 2), a statistical significant correlation could not be established (Table S3). A more diverse pool of strains should be sampled in order to evaluate whether the presence of the toxin encoding genes is a main characteristic of the genomospecies, as was previously suggested (Fagerlund et al. 2004; Carroll et al. 2017; Stevens et al. 2019).

Type / Genomospecies	Number of strains	Cry	Antrax related	Diarrheal toxin	Cereolysin AB	Emetic toxin	Cereolysin O	Cytotoxin K1	Cytotoxin K2	Enterotoxin	Haemolysin BL	Haemolysin II	Immune inhibitor A	Nonhaemolytic Enterotoxin	Phospholipase C
B. wiedmannii FSL W8-0169	124	2%	0%	73%	99%	0%	92%	0%	7%	100%	99%	49%	100%	100%	100%
B. mobilis 0711P9-1	19	5%	0%	16%	100%	0%	74%	0%	11%	100%	58%	58%	100%	100%	89%
B. albus N35-10-2	6	0%	0%	100%	100%	0%	100%	0%	50%	100%	100%	100%	100%	100%	100%
Genomospecies #1	17	0%	0%	100%	100%	0%	100%	0%	0%	100%	100%	100%	100%	100%	100%
Genomospecies #2	6	0%	0%	0%	100%	0%	67%	0%	0%	100%	100%	50%	100%	100%	83%
Genomospecies #3	2	50%	0%	100%	100%	0%	100%	0%	0%	100%	0%	50%	100%	100%	100%
Genomospecies #4	1	0%	0%	0%	100%	0%	100%	0%	0%	100%	100%	0%	100%	100%	100%
Genomospecies #5	1	0%	0%	0%	100%	0%	100%	0%	100%	100%	100%	100%	100%	100%	100%
Genomospecies #6	40	2%	2%	72%	100%	0%	100%	0%	0%	100%	100%	82%	100%	100%	100%
Genomospecies #7	1	0%	0%	0%	100%	0%	100%	0%	0%	100%	100%	100%	100%	100%	100%
Genomospecies #8	5	60%	60%	0%	100%	0%	100%	0%	0%	100%	100%	100%	100%	100%	100%
B. anthracis Ames Ancestor	177	3%	58%	6%	99%	0%	97%	0%	27%	100%	11%	92%	99%	99%	88%
B. paranthracis Mn5	101	3%	2%	10%	100%	24%	100%	0%	24%	100%	2%	2%	100%	99%	99%
B. tropicus N24	61	3%	7%	0%	100%	0%	98%	0%	89%	100%	87%	85%	100%	100%	85%
B. pacificus EB422	38	0%	0%	24%	100%	0%	82%	0%	74%	100%	0%	0%	97%	92%	92%
Genomospecies #9	12	42%	0%	0%	100%	0%	100%	0%	67%	100%	92%	100%	100%	100%	100%
Genomospecies #10	4	25%	0%	0%	100%	0%	100%	0%	100%	100%	100%	75%	100%	100%	100%
Genomospecies #11	1	0%	0%	0%	100%	0%	100%	0%	100%	100%	0%	0%	100%	100%	100%
Genomospecies #12	1	0%	100%	0%	100%	0%	100%	0%	0%	100%	0%	0%	0%	0%	100%
Genomospecies #13	2	0%	0%	0%	100%	0%	100%	0%	0%	100%	100%	100%	100%	100%	100%
Genomospecies #14	2	0%	0%	0%	100%	0%	100%	0%	0%	100%	100%	100%	100%	100%	100%
Genomospecies #15	19	0%	0%	0%	100%	0%	100%	0%	5%	100%	100%	100%	100%	100%	100%
B. cereus ATCC 14579	569	23%	7%	99%	100%	0%	99%	0%	98%	100%	98%	13%	100%	98%	98%
B. thuringiensis ATCC 10792	341	20%	2%	45%	100%	0%	99%	0%	87%	100%	91%	96%	100%	100%	98%
B. toyonensis BCT-7112	228	18%	9%	100%	100%	0%	61%	0%	11%	100%	100%	2%	100%	100%	99%
B. pseudomycoides DSM 12442	67	27%	0%	100%	100%	0%	13%	0%	0%	100%	88%	0%	100%	97%	0%
B. gaemokensis JCM 15801	2	0%	0%	0%	100%	0%	100%	0%	0%	100%	100%	0%	100%	100%	0%
B. manliponensis JCM_15802	1	0%	0%	100%	100%	0%	0%	0%	0%	100%	0%	0%	100%	100%	0%
B. bingmayongensis FJAT_13831	1	0%	0%	0%	100%	0%	0%	0%	0%	100%	0%	0%	100%	100%	0%
Genomospecies #16	1	0%	0%	100%	100%	0%	100%	0%	0%	100%	100%	0%	100%	0%	0%
Genomospecies #17	1	0%	0%	100%	100%	0%	100%	0%	0%	100%	0%	0%	100%	0%	0%
Genomospecies #18	1	0%	0%	100%	100%	0%	100%	0%	0%	100%	0%	0%	100%	0%	0%
Genomospecies #19	8	0%	0%	12%	100%	0%	100%	0%	0%	100%	0%	0%	100%	50%	0%
Genomospecies #20	2	0%	0%	0%	100%	0%	50%	0%	0%	100%	100%	0%	100%	100%	0%
Genomospecies #21	1	0%	0%	0%	100%	0%	100%	0%	0%	100%	100%	0%	100%	100%	0%
Genomospecies #22	43	2%	0%	100%	100%	0%	7%	0%	0%	100%	95%	0%	100%	0%	0%
Genomospecies #23	5	0%	0%	0%	100%	0%	0%	0%	0%	100%	100%	0%	100%	0%	0%
B. mycoides ATCC 6462; *B. weihenstephanensis* NBRC	54	24%	0%	76%	100%	6%	87%	0%	0%	100%	100%	6%	100%	100%	91%
B. nitratireducens 4049	68	0%	3%	62%	100%	0%	100%	0%	0%	100%	99%	0%	100%	99%	0%
B. paramycoides NH24A2	8	0%	0%	50%	100%	0%	75%	0%	0%	100%	88%	0%	100%	100%	100%
B. proteolyticus TD42	3	0%	0%	100%	100%	0%	100%	0%	0%	100%	100%	0%	100%	100%	0%
Genomospecies #24	1	0%	0%	100%	100%	0%	100%	0%	0%	100%	100%	0%	100%	100%	100%
Genomospecies #25	2	0%	0%	0%	100%	0%	100%	0%	0%	100%	100%	0%	100%	100%	0%
Genomospecies #26	1	0%	0%	100%	100%	0%	100%	0%	0%	100%	100%	0%	100%	100%	100%
Genomospecies #27	1	0%	0%	0%	100%	0%	100%	0%	0%	100%	100%	100%	100%	100%	100%
Genomospecies #28	4	0%	0%	100%	100%	0%	100%	0%	0%	100%	100%	0%	100%	100%	100%
Genomospecies #29	1	0%	0%	0%	100%	0%	100%	0%	0%	100%	100%	0%	100%	100%	100%
Genomospecies #30	14	7%	0%	71%	100%	14%	93%	0%	0%	100%	100%	7%	100%	100%	86%
Genomospecies #31	6	50%	0%	0%	83%	0%	83%	0%	0%	100%	100%	0%	100%	100%	0%
Genomospecies #32	3	0%	0%	0%	100%	0%	100%	0%	0%	100%	100%	0%	100%	100%	100%
B. cytotoxicus NVH 391-98	14	0%	0%	100%	100%	0%	0%	100%	0%	100%	0%	0%	93%	100%	0%
B. luti TD41	3	0%	0%	100%	100%	0%	100%	0%	0%	100%	0%	0%	100%	100%	0%
Genomospecies #33	5	0%	0%	0%	100%	0%	100%	0%	0%	100%	100%	0%	100%	100%	0%
Genomospecies #34	6	0%	0%	33%	100%	0%	0%	0%	0%	100%	83%	0%	100%	100%	0%
Genomospecies #35	2	100%	0%	0%	0%	0%	0%	0%	0%	100%	100%	0%	100%	0%	0%
Genomospecies #36	2	0%	50%	100%	100%	0%	100%	0%	100%	100%	100%	100%	100%	100%	100%
Genomospecies #37	7	0%	0%	100%	100%	0%	100%	0%	0%	100%	100%	0%	100%	100%	0%

Clade 1

Clade 2

Other *B. cereus* group strains

100%

0%

In quest of distinguishing features among highly related genomospecies of the *B. cereus* group

In order to shed light on the reported difficulties when defining *B. cereus* group genomospecies, COG categories of all soft-core proteins were determined and used to perform a comparative analysis using PhyloLM. As it is shown in Fig. 3, *B. pseudomycoides*, the related genomospecies #22, and *B. cytotoxicus* showed to be enriched in proteins associated with mobile elements ($p < 0.05$, Fig. 3). The latter group, as well as members of *B. paranthracis, B. toyonensis*, genomospecies #6, and #9, were enriched in functions associated with motility ($p < 0.01$, Fig. 3). *B. thuringiensis* genomospecies members contrasted with *B. cereus* sensu stricto in the increased presence of functions related to lipid transport and metabolism, cell division, and intracellular trafficking and secretion ($p < 0.01$, Fig. 3). On the other hand, members of *B. tropicus* showed to be enriched in the category function of metabolism and transport of nucleotides, coenzyme metabolism, and molecular chaperones and related functions ($p < 0.05$, Fig. 3). Functions related to secondary metabolites biosynthesis, transport and catabolism, biogenesis of cell wall, membrane or envelope, motility, secretion and signal transduction

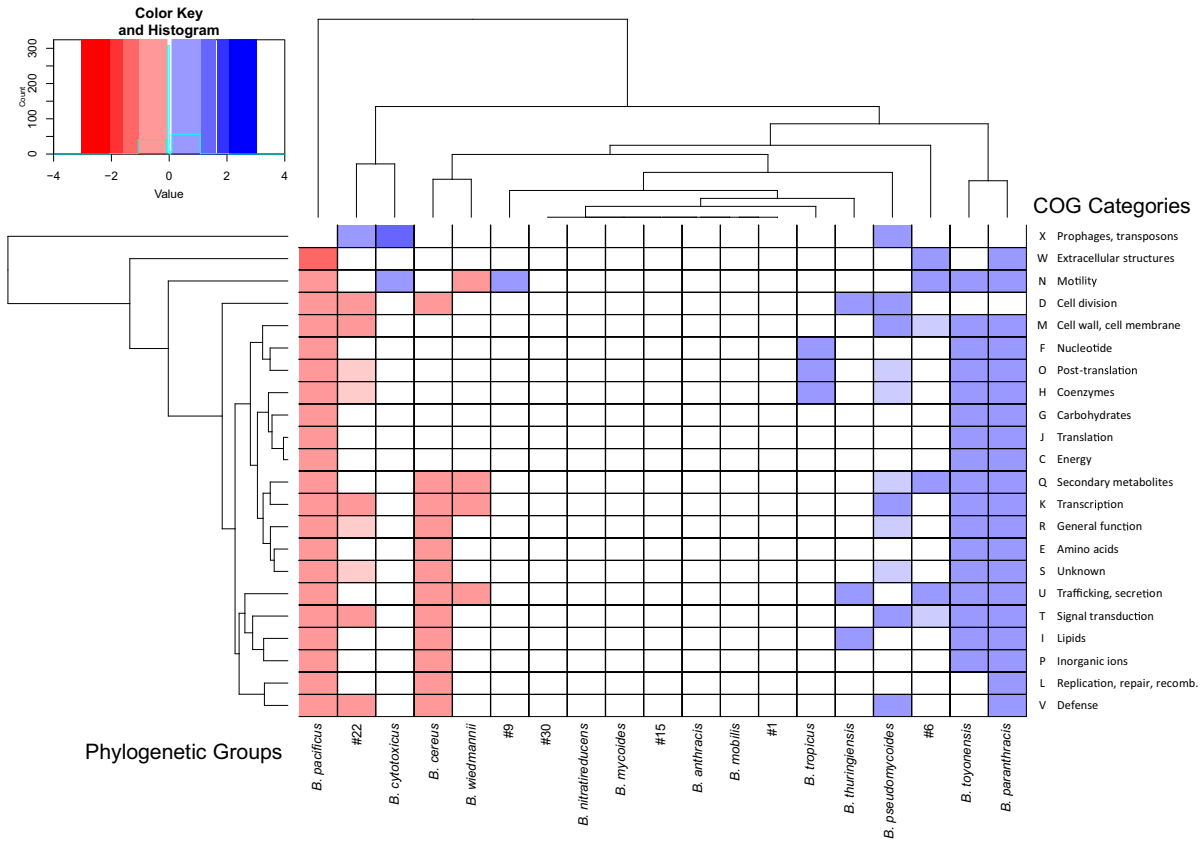

Fig. 3 Difference in COG category abundance. A Normalized Cluster of differences in gene categories between genomes of the same genomospecies. The heat map indicates the level of enrichment or depletion based on a PhyloLM test. Significant cells are shown colored (p value < 0.05, FDR corrected). Hot colored cells indicate significantly more genes. One-letter abbreviations for the functional categories: C, Energy production and conversion; D, Cell division, chromosome partitioning; E, Amino acid transport and metabolism; F, Nucleotide transport and metabolism; G, Carbohydrate transport and metabolism; H, Coenzyme transport and metabolism; I, Lipid transport and metabolism; J, Translation, ribosomal structure and biogenesis; K, Transcription; L, Replication, recombination and repair; M, Cell wall, membrane or envelope biogenesis; N, Cell motility; O, Posttranslational modification, protein turnover, chaperones; P, Inorganic ion transport and metabolism; Q, Secondary metabolites biosynthesis, transport and catabolism; R, General function prediction only; S, Function unknown; T, Signal transduction mechanisms; U, Secretion; V, Defense mechanisms; W, Extracellular structures; X, Mobilome: prophages, transposons

mechanisms were found to be enriched in members of genomospecies #6 ($p < 0.05$, Fig. 3). *B. cereus* sensu stricto members and *B. toyonensis* were observed to be depleted and enriched in many broad function protein categories, respectively ($p < 0.01$, Fig. 3). It was also evident that phylogenetically closed relatives such as members of *B. pseudomycoides* and genomospecies #22; *B. wiedmannii* and genomospecies #6 as well as *B. paranthracis* and *B. pacificus*, showed upweighted profile differences as consequence of the application of a phylogenetic sensitive approach, as was observed in other studies (Bradley et al. 2018).

Next, a search for specific traits differing between Clades 1 and 2 was conducted. In order to achieve this, we selected those genomospecies under analysis with more than 60 available genome sequences. A total of 73 gene differences were identified in this genome subset, using as a benchmark the presence of an ortholog in at least 90% of the genomes on one side of a phylogenetic node, and its simultaneous absence in at least 90% of the genomes on the other side of the node (Fig. 4 and Table S4). Clade 2 strains were enriched in a gene cluster putatively involved in sulfite export and in the bacillolysin proteinase gene ($p = 3.4\ 10^{-6}$). The latter has been reported to be secreted and presumably play a role in *B. cereus* pathogenicity when tested in a *Galleria mellonella* larvae model (Mazzantini et al. 2016). On the other hand, Clade 1 was enriched in gene clusters that encode for a permease ($p < 4.35\ 10^{-14}$) and an intermembrane metalloprotease ($p < 1.75\ 10^{-25}$) of unknown function. We additionally observed that a cluster of genes involved in cell wall or exosporium biosynthesis, encoding several glycosyltransferases, were particularly enriched in *B. cereus* sensu stricto strains with respect to its nearest-neighbor *B. thuringiensis* genomospecies ($p < 1.73\ 10^{-13}$, Table S4). Interestingly, the latter species was enriched in a beta-channel forming cytolysin ($p = 2.93\ 10^{-10}$, Table S4). This supports the finding of Anderson and collaborators (2005) that described a cytolysin with homology to *Streptococcus agalactiae* Cyl, which was present exclusively in *B. thuringiensis* genomospecies. Differences in 117 genes between *B. thuringiensis gv. cytolyticus* and *B. thuringiensis gv. thuringiensis* genomovars were also identified (Fig. 4 and Table S4). Remarkably, the ATPase component (EssC, $p = 1.85\ 10^{-5}$), the membrane proteins EssA and EssB as well as seven putative secretion effector proteins of Type 7 secretion system (T7SS), including three WXG100-family proteins ($p < 2.47\ 10^{-5}$, Table S4) were found to be enriched in *B.*

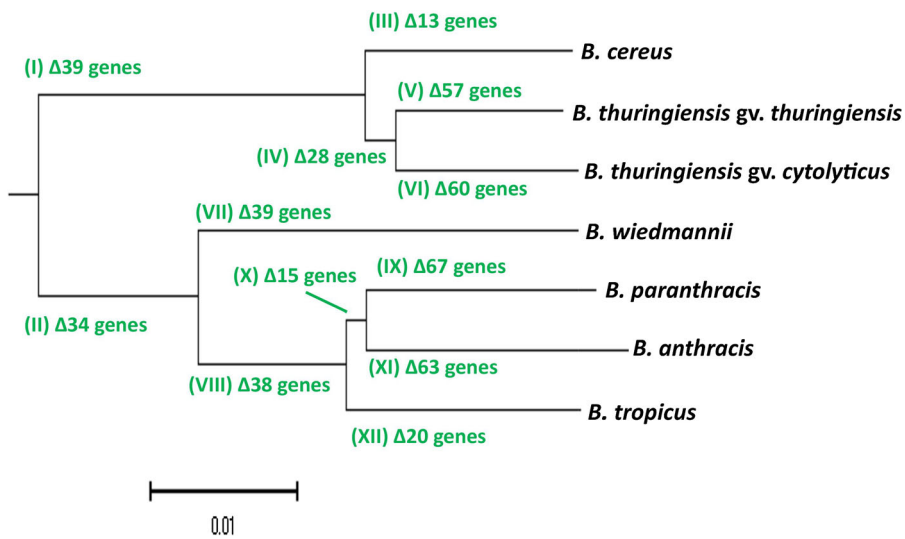

Fig. 4 Changes in genes at each branch point for selected genomospecies of *B. cereus* group. The change in genes signifies the genes that are found in > 90% of the genomes on that side of the branch and < 10% of the genomes on the opposite side. This analysis was generated with the genome comparator function implemented under BIGSdb software (Jolley and Maiden 2010). A complete genome was used as a reference strain for each genomospecies and was used to BLAST all genomes in the set. Specific differences in gene content among *B. cereus* group strains as well as reference strains used in the analysis are detailed in Table S4(I-XII)

thuringiensis gv. cytolyticus. This type of secretion system was identified in Actinobacteria and Firmicutes but not in Gram-negative bacteria (Pallen 2002). Among Clade 1 strains it is noteworthy that a petrobactin biosynthesis cluster was significantly enriched in *B. anthracis* with respect to *B. paranthracis* ($p < 2.18\ \ 10^{-8}$, Table S4). Differences between *B. anthracis* and *B. paranthracis* cell wall carbohydrate biosynthesis clusters were also identified ($p < 0.014$). Additional differences among *B. cereus* group strains are detailed in Table S4(I-XII) in the supplementary information accompanying this article.

Discussion

Members of the *B. cereus* group were ubiquitously distributed in diverse environments and play critical roles in insect pest management, food production and human health (Liu et al. 2017a; Méric et al. 2018). However, the improper use of objective criteria in their classification has resulted in error-prone species circumscriptions (Maughan and Van der Auwera 2011). Here, by means of a core-based MLSA and multiple comparative analysis of ANI values, we observed that Clade 2 members could be circumscribed in *B. cereus* sensu stricto and *B. thuringiensis* of genomospecies. Interestingly, our benchmark analysis allows discerning differences in clusters of conserved glycosyltransferases associated with cell wall or exosporium carbohydrates biosynthesis. These glycosyltransferases likely catalyze unique carbohydrate linkages, since genes for new classes of sugar donors (e.g. sugar nucleotides) are not observed. Differences in bacterial surface structures, which are important targets in serological identifications, have been previously described among *B. cereus* group members (Anderson et al. 2005). Moreover, among the differences identified between the recently defined genomovars of *B. thuringiensis* (Baek et al. 2019), the enrichment of T7SS components in the genomovariety *cytolyticus* should be emphasized. This type of secretion system contributes to virulence and protective immunity in *Mycobacterium tuberculosis* and to the establishment of persistent infections in *Staphylococcus aureus* (Pallen 2002; Garufi et al. 2008). In addition, it was shown that anthrax-infected guinea pigs developed IgG-type antibodies against effectors of the system, suggesting that they are expressed and secreted during the infection (Garufi et al. 2008). However, gene clusters encoding a T7SS were also identified in non-pathogenic bacteria such as *B. subtilis, Streptomyces coelicolor*, and in *M. smegmatis* (Huppert et al. 2014), generating controversies on its actual physiological role. On the other hand, the presence of Cry crystals, key factors of bacterial pathogenicity towards invertebrates, was conventionally considered as a taxonomic marker that defines *B. thuringiensis* strains. Indeed, the correlation between the capability of producing entomotoxins and species designation was thoroughly investigated (Ceuppens et al. 2013; Zheng et al. 2017; Carroll et al. 2017). Nevertheless, "phylogenetic signals" defined as the tendency for related species to resemble each other, were not included in those correlation studies. Noteworthy, it was demonstrated that methods that do not consider phylogenetic signals result in misleadingly low p values leading to false conclusions (Ives and Garland 2014). This is because the main assumptions of standard statistical procedures, such as independence and identical distribution of residuals from a regression model, are not fulfilled at high phylogenetic signals (Felsenstein 1985; Ives and Garland 2010). Recently, phylogeny-aware methods based on linear regression models were applied at wide-genome scale to study the genomes of plant-associated bacteria (Levy et al. 2018) or metagenomic data (Bradley et al. 2018). By using a phylogeny-correcting approach we hereby observed that *cry* genes not only positively correlated with *B. thuringiensis* genomospecies but to all Clade 2 members, including *B. cereus* sensu stricto. Both, the high metabolic burden that imply specialized insecticidal *cry* gene expression (Méric et al. 2018), as well as the high degree of sexual isolation (Patiño-Navarrete and Sanchis 2017), may explain the enrichment of *cry* genes in Clade 2. In addition, a strong positive correlation ($p = 7.3\ \ 10^{-08}$) of Cry-encoding genes with strains isolated from insects was observed, indicating that the capability to produce crystals is more related to niche adaptation rather than to strains phylogenetic relationships. Commercial biopesticide formulations based on *B. thuringiensis* represent a worldwide for-profit industry (Maughan and Van der Auwera 2011). Hence, our results are in agreement with the recommendation of EFSA regarding the application of whole-genome sequencing to provide unambiguous identification of strains used as biopesticides as a prerequisite for its safety evaluation (EFSA

BIOHAZ Panel 2016). In opposition, the recent taxonomic framework proposed by Carroll et al. suggests that Clade 2 members should be clustered into the *B. cereus* species (2020). In an attempt to reconcile phylogenomics relationships and phenotypic traits that can be lost or gained by particular strains within a species or across multiple species, the authors make use of subspecies names and biovar terms to account for the heterogeneity of clinically and industrially important phenotypes (Carroll et al. 2020). The authors applied biovar Thuringiensis to all Cry, Cyt, or Vip toxin-producing isolates. Following this criterion the biovar Thuringiensis would be assigned to members of 5 different *B. cereus* group species ("*B. mosaicus", B. mycoides, B. cereus, B. toyonensis,* and *B. pseudomycoides*).

Species assignation, implicitly or not, is used to predict strains behavior or performance (Gevers et al. 2005; Torres Manno et al. 2018) and therefore favor or discourage commercial initiatives or more exhaustive safety safeguards regarding the concern or usage of bacterial isolates. The current drawback of this inherent use of taxonomic information is that many species are defined by either single or over-represented isolates (Rasko et al. 2005; Felis and Dellaglio 2007). Poor diversity represented in species description usually is due to the fact that bacterial sampling, strain selection, and data publication are not set at random, but instead have anthropocentric biases (based on human health relevance, economical profits, scientific impact, etc.). Such biases in species definition drastically affect the accuracy and feasibility of comparative studies. We found that the concept of genomospecies or genomovars as defined by Baek et al. (2019) could help to better describe the diversity of complex taxonomic groups such as *B. cereus.* For instance, our genome-scale approaches suggest that the clonal emetic *B. cereus* strains linked to outbreaks should be reassigned as members of *B. paranthracis* genomospecies, supporting the proposal of Carroll et al. (2019). In this regard, we have observed that the emetic toxin genes show no positive correlation with *B. paranthracis* genomospecies when phylogenetic signals are considered. This is due the fact that genes involved in cereulide biosynthetic are plasmid-encoded (Ehling-Schulz et al. 2006). Moreover, diversity of emetic-producing *B. cereus sensus lato* were previously reported (Apetroaie et al. 2005). On the other hand, we observed that anthrax toxin and capsule genes encoded by pXO1 and pXO2 plasmids positively correlated with *B. anthracis* species even in a phylogenetic-corrected analysis. Nevertheless, our results indicated that some strains of *B. anthracis* genomospecies, submitted at NCBI as *B. cereus,* should be renamed as *B. anthracis* since they could not be distinguished from *B. anthracis* Ames at genomic level, as was previously suggested (Maughan and Van der Auwera 2011). In opposition, other authors have proposed that these strains should be designated as *B. anthracis* sensu lato or *B. cereus var anthracis* (Okinaka et al. 2006; Kolstø et al. 2009). *B. anthracis*-associated genes such as *capABCDE* and/or *hasA* were also positively correlated with other species such as *B. toyonensis*. Remarkably, the archetype strain of this species is the non-toxigenic strain *B. toyonensis* BCT-7112^{T}, used as probiotics in animal nutrition in a wide range of countries including Europe and Japan (Jiménez et al. 2013). The isolation of *B. toyonensis* strains from environmental samples encoding pXO1- or pXO2-like plasmids was previously reported (Van der Auwera et al. 2013). Diarrheal toxin and phospholipase C encoding genes also positively correlated with *B. toyonensis* species in comparison with other related species (Table S3). In addition, while strains belonging to *B. toyonensis* were mainly isolated from the environmental or plant sources, *B. toyonensis* FDAARGOS_235 and BacAer strains were isolated from blood and epithelial infections, and are thus considered as opportunistic pathogens. It is noteworthy that common pathogenic determinants or traits could have originated from evolutionary convergence or horizontal gene transfer and therefore may not accurately reflect evolutionary relatedness among strains (Maughan and Van der Auwera 2011).

In order to resolve nomenclatural ambiguities, Carroll et al. proposed clustering all members of *B. paranthracis, B. anthracis* (clonal or not) as well as other less related species such as *B. albus*, *B. mobilis*, *B. pacificus*, *B. tropicus*, and *B. wiedmannii* into the *"B. mosaicus"* species (2020). In addition, the authors suggested that those strains possessing the anthrax toxin-encoding genes *cya, lef,* and *pagA* be assigned to the biovar Anthracis, whereas those encoding the *cesABCD* genes be assigned to the biovar Emeticus. Hence, strains that are not tightly phylogenomically related to *B. anthracis* but possess the transmissible anthrax toxin-encoding genes should be named *B.*

Anthracis. On the other hand, those strains sharing ≥ 99.9 ANI with *B. anthracis* Ames should be referred as "*B. mosaicus* subsp. *anthracis"* or merely "*B. anthracis*". Hence, authors maintain historical congruence supporting its standardized framework in phenotypic or transmissible genetic traits. In this context, our proposed genomospecies boundaries could coexist with Carroll et al. taxonomic framework by referring them as subspecies or genomovarieties. For example, strains of "*B.* Emeticus" (as was suggested by Carroll et al.) could be named as "*B. mosaicus subsp. paranthracis* biovar Emeticus" or "*B. mycoides* biovar Emeticus" in case they were identified as *B. paranthracis* or *B. mycoides* genomospecies members in our approach, respectively.

In our comparative analysis, the enrichment of traits related to the regulation of gene expression, spore coat, exopolysaccharide biosynthesis, and membrane transport among *B. cereus* group bacteria has been pinpointed (Table S4). These findings are in concordance with previous studies (Anderson et al. 2005) and greatly expand our understanding of the genetic differences amongst these important species. Nevertheless, it is noteworthy that our approach has distinguished for the first time the petrobactin gene cluster as an important marker between the phylogenetically related *B. anthracis* and *B. paranthracis* genomospecies. Remarkably, this siderophore was shown to be involved in iron acquisition, protection against oxidative stress, and plays a vital role in the many stages of *B. anthracis* infection, including macrophage survival, bloodstream growth and facilitating transmission between mammalian hosts (Hagan et al. 2018). Hence, our analysis may contribute to refocus current studies aiming to broad the comprehension of *B. anthracis* infections. Nonetheless, the necessity of having a confident method for genomospecies phylogenetic delimitation results evident, in order to perform comparative studies among related strains. In this work, we aimed to generate a conceptual framework in which the genomospecies of the *B. cereus* group were accurately defined. As a result, we propose the inclusion of 37 new genomospecies as well as the re-assignation of 832 strains. We expect our framework to be adopted by the scientific community as a valuable guide for future epidemiological analysis as well as population dynamics and evolution studies.

Acknowledgements We would like to thank Agencia Nacional de Promoción Científica y Tecnológica for financial support. MATM is CONICET fellow; CM, GDR, and ME are researchers of the same institution. CAD is a researcher at USDA. We would also like to thank Prunello M. from UNR for assistance with statistical analysis of the results that greatly improved the manuscript. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the U.S. Department of Agriculture. The mention of firm names or trade products does not imply that they are endorsed or recommended by the USDA over other firms or similar products not mentioned. USDA is an equal opportunity provider and employer.

Author contribution All authors contributed conception and design of the study; MATM and CAD performed the in silico analyses; MATM performed the statistical analysis; All authors interpreted the evolutionary relationships; ME wrote the first draft of the manuscript; All authors wrote sections of the manuscript, contributed to manuscript revision, read and approved the submitted version.

Funding information This study was funded by Agencia Nacional de Promoción Científica y Tecnológica PICT 2018-01872 to ME, PICT2017-3536, PIP2017-11220170100377CO to GDR, and Agricultural Research Service Project Number 5010-22410-019-00-D to CAD.

Availability of data and materials All data analyzed during the current study were downloaded from public databases (NCBI), and dates of download are provided in the text. A list of all assembly accessions used in this study is provided in Supplementary Table S1. ANI values of the 4.665.600 genome comparisons, all alignments as well as their corresponding phylogenetic trees in Newick format are available in the GitHub repository (https://github.com/torresmanno/Bacillus-cereus-data).

Compliance with ethical standards

Conflict of interest The authors (MATM, GDR, CM, CAD, and ME) have no conflicts of interest to declare.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Alcaraz L, Moreno-Hagelsieb G, Eguiarte LE, Souza V, Herrera-Estrella L, Olmedo G (2010) Understanding the evolutionary relationships and major traits of *Bacillus* through comparative genomics. BMC Genom 11:332. https://doi.org/10.1186/1471-2164-11-332

Amor MGB, Siala M, Zayani M, Grosset N, Smaoui S, Messadi-Akrout F, Baron F, Jan S, Gautier M, Gdoura R (2018)

Isolation, identification, prevalence, and genetic diversity of *Bacillus cereus* group bacteria from different foodstuffs in Tunisia. Front Microbiol 9:1–12. https://doi.org/10.3389/fmicb.2018.00447

Anderson I, Sorokin A, Kapatral V, Reznik G, Bhattacharya A, Mikhailova N, Burd H, Joukov V, Kaznadzey D, Walunas T, D'Souza M, Larsen N, Pusch G, Liolios K, Grechkin Y, Lapidus A, Goltsman E, Chu L, Fonstein M, Ehrlich SD, Overbeek R, Kyrpides N, Ivanova N (2005) Comparative genome analysis of *Bacillus cereus* group genomes with *Bacillus subtilis*. FEMS Microbiol Lett 250:175–184. https://doi.org/10.1016/j.femsle.2005.07.008

Apetroaie C, Andersson MA, Spröer C, Tsitko I, Shaheen R, Jääskeläinen EL, Wijnands LM, Heikkilä R, Salkinoja-Salonen MS (2005) Cereulide-producing strains of *Bacillus cereus* show diversity. Arch Microbiol 184:141–151. https://doi.org/10.1007/s00203-005-0032-1

Baek I, Lee K, Goodfellow M, Chun J (2019) Comparative genomic and phylogenomic analyses clarify relationships within and between *Bacillus cereus* and *Bacillus thuringiensis*: proposal for the recognition of two *Bacillus thuringiensis* Genomovars. Front Microbiol 10:1–11. https://doi.org/10.3389/fmicb.2019.01978

Bazinet AL (2017) Pan-genome and phylogeny of *Bacillus cereus* sensu *lato*. BMC Evol Biol 17:1–16. https://doi.org/10.1186/s12862-017-1020-1

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Beno SM, Orsi RH, Cheng RA, Kent DJ, Kovac J, Duncan DR, Martin NH, Wiedmann M (2019) Genes associated with psychrotolerant *Bacillus cereus* group isolates. Front Microbiol 10:1–14. https://doi.org/10.3389/fmicb.2019.00662

Bhandari V, Ahmod NZ, Shah HN, Gupta RS (2013) Molecular signatures for *Bacillus* species: Demarcation of the *Bacillus subtilis* and *Bacillus cereus* clades in molecular terms and proposal to limit the placement of new species into the genus *Bacillus*. Int J Syst Evol Microbiol 63:2712–2726. https://doi.org/10.1099/ijs.0.048488-0

Borowiec ML (2016) AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ 4:e1660. https://doi.org/10.7717/peerj.1660

Bradley PH, Nayfach S, Pollard KS (2018) Phylogeny-corrected identification of microbial gene families relevant to human gut colonization

Carlin F, Brillard J, Broussolle V, Clavel T, Duport C, Jobin M, Guinebretière MH, Auger S, Sorokine A, Nguyen-Thé C (2010) Adaptation of *Bacillus cereus*, an ubiquitous worldwide-distributed foodborne pathogen, to a changing environment. Food Res Int 43:1885–1894. https://doi.org/10.1016/j.foodres.2009.10.024

Carroll LM, Kovac J, Miller RA, Wiedmann M (2017) Rapid, high-throughput identification of anthrax-causing and emetic *Bacillus cereus* group genome assemblies via BTyper, a computational tool for virulence-based classification of *Bacillus cereus* group isolates by using nucleotide sequencing. Appl Environ Microbiol 83:1–19. https://doi.org/10.1128/aem.01096-17

Carroll LM, Wiedmann M, Mukherjee M, Nicholas DC, Mingle LA, Dumas NB, Cole JA, Kovac J (2019) Characterization of emetic and diarrheal *Bacillus cereus* strains from a 2016 foodborne outbreak using whole-genome sequencing: addressing the microbiological, epidemiological, and bioinformatic challenges. Front Microbiol. https://doi.org/10.3389/fmicb.2019.00144

Carroll LM, Wiedmann M, Kovac J (2020) Proposal of a taxonomic nomenclature for the *Bacillus cereus* group which reconciles genomic definitions of bacterial species with clinical and industrial phenotypes. MBio 11:1–15. https://doi.org/10.1128/mBio.00034-20

Ceuppens S, Boon N, Uyttendaele M (2013) Diversity of *Bacillus cereus* group strains is reflected in their broad range of pathogenicity and diverse ecological lifestyles. FEMS Microbiol Ecol 84:433–450. https://doi.org/10.1111/1574-6941.12110

Chun J, Oren A, Ventosa A, Christensen H, Arahal DR, da Costa MS, Rooney AP, Yi H, Xu XW, De Meyer S, Trujillo ME (2018) Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. Int J Syst Evol Microbiol 68:461–466. https://doi.org/10.1099/ijsem.0.002516

Ciufo S, Kannan S, Sharma S, Badretdin A, Clark K, Turner S, Brover S, Schoch CL, Kimchi A, DiCuccio M (2018) Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. Int J Syst Evol Microbiol 68:2386–2392. https://doi.org/10.1099/ijsem.0.002809

Didelot X, Barker M, Falush D, Priest FG (2009) Evolution of pathogenicity in the *Bacillus cereus* group. Syst Appl Microbiol 32:81–90. https://doi.org/10.1016/j.syapm.2009.01.001

EFSA BIOHAZ Panel (2016) Risks for public health related to the presence of *Bacillus cereus* and other *Bacillus spp.* including *Bacillus thuringiensis* in foodstuffs EFSA panel on biological hazards (BIOHAZ). https://doi.org/10.2903/j.efsa.2016.4524

Ehling-Schulz M, Fricker M, Grallert H, Rieck P, Wagner M, Scherer S (2006) Cereulide synthetase gene cluster from emetic *Bacillus cereus*: Structure and location on a mega virulence plasmid related to *Bacillus anthracis* toxin plasmid pXO1. BMC Microbiol 6:1–11. https://doi.org/10.1186/1471-2180-6-20

Espariz M, Zuljan FA, Esteban L, Magni C (2016) Taxonomic Identity resolution of highly phylogenetically related strains and selection of phylogenetic markers by using genome-scale methods: the *Bacillus pumilus* group case. PLoS ONE 11:e0163098. https://doi.org/10.1371/journal.pone.0163098

Fagerlund A, Ween O, Lund T, Hardy SP, Granum PE (2004) Genetic and functional analysis of the *cytK* family of genes in *Bacillus cereus*. Microbiology. https://doi.org/10.1099/mic.0.26975-0

Felis GE, Dellaglio F (2007) On species descriptions based on a single strain: proposal to introduce the status species proponenda (sp. pr.). Int J Syst Evol Microbiol 57:2185–2187. https://doi.org/10.1099/ijs.0.64931-0

Felsenstein J (1985) Phylogenies and the comparative method. Am Nat 125:1–15. https://doi.org/10.1086/284325

Garufi G, Butler E, Missiakas D (2008) ESAT-6-like protein secretion in *Bacillus anthracis*. J Bacteriol 190:7004–7011. https://doi.org/10.1128/JB.00458-08

Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Van de Peer Y, Vandamme P, Thompson FL, Swings J (2005) Re-evaluating prokaryotic species. Nat Rev Microbiol 3:733–739. https://doi.org/10.1038/nrmicro1236

Guérin A, Rønning HT, Dargaignaratz C, Clavel T, Broussolle V, Mahillon J, Granum PE, Nguyen-The C (2017) Cereulide production by *Bacillus weihenstephanensis* strains during growth at different pH values and temperatures. Food Microbiol. https://doi.org/10.1016/j.fm.2017.02.006

Guinebretière MH, Thompson FL, Sorokin A, Normand P, Dawyndt P, Ehling-Schulz M, Svensson B, Sanchis V, Nguyen-The C, Heyndrickx M, De Vos P (2008) Ecological diversification in the *Bacillus cereus* Group. Environ Microbiol 10:851–865. https://doi.org/10.1111/j.1462-2920.2007.01495.x

Guinebretière MH, Velge P, Couvert O, Carlin F, Debuyser ML, Nguyen-The C (2010) Ability of *Bacillus cereus* group strains to cause food poisoning varies according to phylogenetic affiliation (groups I to VII) rather than species affiliation. J Clin Microbiol 48:3388–3391. https://doi.org/10.1128/JCM.00921-10

Guinebretière MH, Auger S, Galleron N, Contzen M, de Sarrau B, de Buyser ML, Lamberet G, Fagerlund A, Granum PE, Lereclus D, de Vos P, Nguyen-The C, Sorokin A (2013) *Bacillus cytotoxicus sp. nov.* is a novel thermotolerant species of the *Bacillus cereus* group occasionally associated with food poisoning. Int J Syst Evol Microbiol 63:31–40. https://doi.org/10.1099/ijs.0.030627-0

Hagan AK, Plotnick YM, Dingle RE, Mendel ZI, Cendrowski SR, Sherman DH, Tripathi A, Hanna PC (2018) Petrobactin protects against oxidative stress and enhances sporulation efficiency in *Bacillus anthracis* sterne. MBio 9:1–14. https://doi.org/10.1128/mBio.02079-18

Han CS, Xie G, Challacombe JF, Altherr MR, Bhotika SS, Bruce D, Campbell CS, Campbell ML, Chen J, Chertkov O, Cleland C, Dimitrijevic M, Doggett NA, Fawcett JJ, Glavina T, Goodwin LA, Hill KK, Hitchcock P, Jackson PJ, Keim P, Kewalramani AR, Longmire J, Lucas S, Malfatti S, Mcmurry K, Meincke LJ, Misra M, Moseman BL, Mundt M, Munk AC, Okinaka RT, Reilly LP, Richardson P, Robinson DL, Rubin E, Saunders E, Tapia R, Tesmer JG, Thayer N, Thompson LS, Tice H, Ticknor LO, Wills PL, Brettin TS, Gilna P (2006) Pathogenomic sequence analysis of *Bacillus cereus* and *Bacillus thuringiensis* ISOLATES CLOSELY RELATED to *Bacillus anthracis*†. 188:3382–3390. https://doi.org/10.1128/JB.188.9.3382

Hong HA, Le HD, Cutting SM (2005) The use of bacterial spore formers as probiotics. FEMS Microbiol Rev 29:813–835. https://doi.org/10.1016/j.femsre.2004.12.001

Huppert LA, Ramsdell TL, Chase MR, Sarracino DA, Fortune SM, Burton BM (2014) The ESX system in *Bacillus subtilis* mediates protein secretion. PLoS ONE. https://doi.org/10.1371/journal.pone.0096267

Ives AR, Garland T (2010) Phylogenetic logistic regression for binary dependent variables. Syst Biol 59:9–26. https://doi.org/10.1093/sysbio/syp074

Ives AR, Garland T (2014) Phylogenetic regression for binary dependent variables. Mod Phylogenetic Comp Methods Appl Evol Biol 59:231–261. https://doi.org/10.1007/978-3-662-43550-2_9

Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S (2018) High throughput ANI analysis of 90 K prokaryotic genomes reveals clear species boundaries. Nat Commun. https://doi.org/10.1038/s41467-018-07641-9

Jiménez G, Urdiain M, Cifuentes A, López-López A, Blanch AR, Tamames J, Kämpfer P, Kolstø AB, Ramón D, Martínez JF, Codoñer FM, Rosselló-Móra R (2013) Description of *Bacillus toyonensis sp. nov.*, a novel species of the *Bacillus cereus* group, and pairwise genome comparisons of the species of the group by means of ANI calculations. Syst Appl Microbiol 36:383–391. https://doi.org/10.1016/j.syapm.2013.04.008

Johler S, Kalbhenn EM, Heini N, Brodmann P, Gautsch S, Bagcioglu M, Contzen M, Stephan R, Ehling-Schulz M (2018) Enterotoxin production of *Bacillus thuringiensis* isolates from biopesticides, foods, and outbreaks. Front Microbiol 9:1–11. https://doi.org/10.3389/fmicb.2018.01915

Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL (2008) NCBI BLAST: a better web interface. Nucleic Acids Res 36:W5–W9. https://doi.org/10.1093/nar/gkn201

Jolley KA, Maiden MCJ (2010) BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-11-595

Jung MY, Paek WK, Park IS, Han JR, Sin Y, Paek J, Rhee MS, Kim H, Song HS, Chang YH (2010) *Bacillus gaemokensis sp. nov.*, isolated from foreshore tidal flat sediment from the Yellow Sea. J Microbiol 48:867–871. https://doi.org/10.1007/s12275-010-0148-0

Kolstø A-B, Tourasse NJ, Økstad OA (2009) What sets *Bacillus anthracis* apart from other *Bacillus* species? Annu Rev Microbiol 63:451–476. https://doi.org/10.1146/annurev.micro.091208.073255

Lazarte JN, Lopez RP, Ghiringhelli PD, Berón CM (2018) *Bacillus wiedmannii* biovar *thuringiensis*: a specialized mosquitocidal pathogen with plasmids from diverse origins. Genome Biol Evol 10:2823–2833. https://doi.org/10.1093/gbe/evy211

Lee I, Kim YO, Park SC, Chun J (2016) OrthoANI: An improved algorithm and software for calculating average nucleotide identity. Int J Syst Evol Microbiol 66:1100–1103. https://doi.org/10.1099/ijsem.0.000760

Letunic I, Bork P (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res 39:W475–W478. https://doi.org/10.1093/nar/gkr201

Levy A, Salas Gonzalez I, Mittelviefhaus M, Clingenpeel S, Herrera Paredes S, Miao J, Wang K, Devescovi G, Stillman K, Monteiro F, Rangel Alvarez B, Lundberg DS, Lu T-Y, Lebeis S, Jin Z, McDonald M, Klein AP, Feltcher ME, Rio TG, Grant SR, Doty SL, Ley RE, Zhao B, Venturi V, Pelletier DA, Vorholt JA, Tringe SG, Woyke T, Dangl JL (2018) Genomic features of bacterial adaptation to plants. Nat Genet 50:138–150. https://doi.org/10.1038/s41588-017-0012-9

Liu B, Liu GH, Hu GP, Cetin S, Lin NQ, Tang JY, Tang WQ, Lin YZ (2014) *Bacillus bingmayongensis sp. nov.*, isolated from the pit soil of Emperor Qin's Terra-cotta warriors in China. Antonie van Leeuwenhoek Int J Gen Mol Microbiol 105:501–510. https://doi.org/10.1007/s10482-013-0102-3

Liu Y, Lai Q, Göker M, Meier-Kolthoff JP, Wang M, Sun Y, Wang L, Shao Z (2015) Genomic insights into the taxonomic status of the *Bacillus cereus* group. Sci Rep 5:1–11. https://doi.org/10.1038/srep14082

Liu Y, Du J, Lai Q, Zeng R, Ye D, Xu J, Shao Z (2017a) Proposal of nine novel species of the *Bacillus cereus* group. Int J Syst Evol Microbiol 67:2499–2508. https://doi.org/10.1099/ijsem.0.001821

Liu Y, Lai Q, Du J, Shao Z (2017b) Genetic diversity and population structure of the *Bacillus cereus* group bacteria from diverse marine environments. Sci Rep 7:1–11. https://doi.org/10.1038/s41598-017-00817-1

Mandic-mulec I, Stefanic P, Elsas JANDVAN (2015) Ecology of Bacillaceae. Bact Spore Mol Syst. https://doi.org/10.1128/microbiolspec.TBS-0017-2013

Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Geer LY, Bryant SH (2017) CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. https://doi.org/10.1093/nar/gkw1129

Maughan H, Van der Auwera G (2011) *Bacillus* taxonomy in the genomic era finds phenotypes to be essential though often misleading. Infect Genet Evol 11:789–797. https://doi.org/10.1016/j.meegid.2011.02.001

Mazzantini D, Celandroni F, Salvetti S, Gueye SA, Lupetti A, Senesi S, Ghelardi E (2016) FlhF is required for swarming motility and full pathogenicity of *Bacillus cereus*. Front Microbiol 7:1–9. https://doi.org/10.3389/fmicb.2016.01644

Méric G, Mageiros L, Pascoe B, Woodcock DJ, Mourkas E, Lamble S, Bowden R, Jolley KA, Raymond B, Sheppard SK (2018) Lineage-specific plasmid acquisition and the evolution of specialized pathogens in *Bacillus thuringiensis* and the *Bacillus cereus* group. Mol Ecol 27:1524–1540. https://doi.org/10.1111/mec.14546

Miller RA, Beno SM, Kent DJ, Carroll LM, Martin NH, Boor KJ, Kovac J (2016) *Bacillus wiedmannii sp. nov.*, a psychrotolerant and cytotoxic *Bacillus cereus* group species isolated from dairy foods and dairy environments. Int J Syst Evol Microbiol 66:4744–4753. https://doi.org/10.1099/ijsem.0.001421

Miller RA, Jian J, Beno SM, Wiedmann M, Kovac J (2018) Intraclade variability in toxin production and cytotoxicity of *Bacillus cereus* group type strains and dairy-associated isolates. Appl Environ Microbiol 84:1–15. https://doi.org/10.1128/AEM.02479-17

Okinaka R, Pearson T, Keim P (2006) Anthrax, but Not *Bacillus anthracis*? PLoS Pathog 2:e122. https://doi.org/10.1371/journal.ppat.0020122

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J (2015) Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics. https://doi.org/10.1093/bioinformatics/btv421

Pallen MJ (2002) The ESAT-6/WXG100 superfamily—and a new Gram-positive secretion system? Trends Microbiol 10:209–212. https://doi.org/10.1016/S0966-842X(02)02345-4

Patiño-Navarrete R, Sanchis V (2017) Evolutionary processes and environmental factors underlying the genetic diversity and lifestyles of *Bacillus cereus* group bacteria. Res Microbiol 168:309–318. https://doi.org/10.1016/j.resmic.2016.07.002

Rasko DA, Altherr MR, Han CS, Ravel J (2005) Genomics of the *Bacillus cereus* group of organisms. FEMS Microbiol Rev 29:303–329

Raymond B, Bonsall MB (2013) Cooperation and the evolutionary ecology of bacterial virulence: The *Bacillus cereus* group as a novel study system. BioEssays 35:706–716. https://doi.org/10.1002/bies.201300028

Rong X, Huang Y (2014) Multi-locus sequence analysis: Taking prokaryotic systematics to the next level. In: Methods in microbiology, 1st edn. Elsevier, Amsterdam, pp 221–251

Rooney AP, Price NPJ, Ehrhardt C, Sewzey JL, Bannan JD (2009) Phylogeny and molecular taxonomy of the *Bacillus subtilis* species complex and description of *Bacillus subtilis* subsp. *inaquosorum* subsp. nov. Int J Syst Evol Microbiol 59:2429–2436. https://doi.org/10.1099/ijs.0.009126-0

Rosselló-Móra R (2012) Towards a taxonomy of Bacteria and Archaea based on interactive and cumulative data repositories. Environ Microbiol 14:318–334. https://doi.org/10.1111/j.1462-2920.2011.02599.x

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics. https://doi.org/10.1093/bioinformatics/btu153

Shapiro BJ, Leducq J-B, Mallet J (2016) What is speciation? PLoS Genet 12:e1005860. https://doi.org/10.1371/journal.pgen.1005860

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. https://doi.org/10.1038/msb.2011.75

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stenfors Arnesen LP, Fagerlund A, Granum PE (2008) From soil to gut: *Bacillus cereus* and its food poisoning toxins. FEMS Microbiol Rev 32:579–606. https://doi.org/10.1111/j.1574-6976.2008.00112.x

Stevens MJA, Tasara T, Klumpp J, Stephan R, Ehling-Schulz M, Johler S (2019) Whole-genome-based phylogeny of *Bacillus cytotoxicus* reveals different clades within the species and provides clues on ecology and evolution. Sci Rep. https://doi.org/10.1038/s41598-018-36254-x

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564–577. https://doi.org/10.1080/10635150701472164

Torres Manno M, Zuljan F, Alarcón S, Esteban L, Blancato V, Espariz M, Magni C (2018) Genetic and phenotypic features defining industrial relevant *Lactococcus lactis, L.*

cremoris and *L. lactis* biovar. diacetylactis strains. J Biotechnol 282:25–31. https://doi.org/10.1016/j.jbiotec.2018.06.345

Torres Manno MA, Pizarro MD, Prunello M, Magni C, Daurelio LD, Espariz M (2019) GeM-Pro: a tool for genome functional mining and microbial profiling. Appl Microbiol Biotechnol 103:3123–3134. https://doi.org/10.1007/s00253-019-09648-8

Van der Auwera GA, Feldgarden M, Kolter R, Mahillon J (2013) Whole-Genome Sequences of 94 Environmental Isolates of *Bacillus cereus* sensu *lato*. Genome Announc 1:1–4. https://doi.org/10.1128/genomeA.00380-13

Whitman WB (2014) The need for change: embracing the genome. In: Methods in microbiology, 1st edn. Elsevier, Amsterdam, pp 1–12

Wu H, Gu Q, Xie Y, Lou Z, Xue P, Fang L, Yu C, Jia D, Huang G, Zhu B, Schneider A, Blom J, Lasch P, Borriss R, Gao X (2019) Cold-adapted Bacilli isolated from the Qinghai-Tibetan Plateau are able to promote plant growth in extreme environments. Environ Microbiol 21:3505–3526. https://doi.org/10.1111/1462-2920.14722

Yoon SH, Ha SM, Kwon S, Lim J, Kim Y, Seo H, Chun J (2017) Introducing EzBioCloud: A taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. Int J Syst Evol Microbiol 67:1613–1617. https://doi.org/10.1099/ijsem.0.001755

Zheng J, Gao Q, Liu L, Liu H, Wang Y, Peng D, Ruan L, Raymond B, Sun M (2017) Comparative genomics of *Bacillus thuringiensis* reveals a path to specialized exploitation of multiple invertebrate hosts. MBio 8:1–14. https://doi.org/10.1128/mBio.00822-17

Zwick ME, Joseph SJ, Didelot X, Chen PE, Bishop-Lilly KA, Stewart AC, Willner K, Nolan N, Lentz S, Thomason MK, Sozhamannan S, Mateczun AJ, Du L, Read TD (2012) Genomic characterization of the *Bacillus cereus* sensu lato species: Backdrop to the evolution of *Bacillus anthracis*. Genome Res 22:1512–1524. https://doi.org/10.1101/gr.134437.111