Research in Microbiology 174 (2023) 104050

Contents lists available at ScienceDirect

Research in Microbiology

journal homepage: www.elsevier.com/locate/resmic

Core genome multilocus sequence typing scheme for *Bacillus cereus* group bacteria

Nicolas J. Tourasse [a, b, *], Keith A. Jolley [c], Anne-Brit Kolstø [a], Ole Andreas Økstad [a, *]

[a] Department of Pharmacology and Pharmaceutical Biosciences, University of Oslo, Norway
[b] University of Bordeaux, CNRS, INSERM, ARNA, UMR 5320, U1212, F-33000 Bordeaux, France
[c] Department of Biology, University of Oxford, UK

ARTICLE INFO

Article history:
Received 31 October 2022
Accepted 28 February 2023
Available online 8 March 2023

Keywords:
Bacillus cereus group
cgMLST
Core genome
Phylogeny
chewBBACA
PubMLST

ABSTRACT

Core genome multilocus sequence typing (cgMLST) employs a strategy where the set of orthologous genes common to all members of a group of organisms are used for phylogenetic analysis of the group members. The *Bacillus cereus* group consists of species with pathogenicity towards insect species as well as warm-blooded animals including humans. While *B. cereus* is an opportunistic pathogen linked to a range of human disease conditions, including emesis and diarrhoea, *Bacillus thuringiensis* is an entomopathogenic species with toxicity toward insect larvae, and therefore used as a biological pesticide worldwide. *Bacillus anthracis* is a classical obligate pathogen causing anthrax, an acute lethal condition in herbivores as well as humans, and which is endemic in many parts of the world. The group also includes a range of additional species, and *B. cereus* group bacteria have been subject to analysis with a wide variety of phylogenetic typing systems. Here we present, based on analyses of 173 complete genomes from *B. cereus* group species available in public databases, the identification of a set of 1568 core genes which were used to create a core genome multilocus typing scheme for the group which is implemented in the PubMLST system as an open online database freely available to the community. The new cgMLST system provides unprecedented resolution over existing phylogenetic analysis schemes covering the *B. cereus* group.

1. Introduction

The *Bacillus cereus* group is a group of closely related bacteria, which includes important pathogens to humans and animals as well as insects. Classically, the *B. cereus* group population (*B. cereus sensu lato*), as defined by Guinebretiere and colleagues, consists of seven core species (*B. cereus* (*sensu stricto*), *Bacillus anthracis*, *Bacillus thuringiensis*, *Bacillus mycoides*, *Bacillus pseudomycoides*, *Bacillus weihenstephanensis*, and *Bacillus cytotoxicus*), phylogenetically divided into seven main clusters and three large clades [1–3].

B. anthracis is the cause of anthrax disease in warm-blooded animals and humans, and is endemic or hyper-endemic in many regions of the world, covering all continents [4]. It has classically been defined as a separate species within the *B. cereus* group based on its ability to cause anthrax disease, a highly serious infectious disease primarily affecting domestic and wild animals around the world, as well as humans. Isolates within the *B. cereus* group able to cause anthrax-like disease have been identified as belonging to phylogenetic clusters outside the *B. anthracis sensu stricto* cluster, causing initial taxonomic confusion, however are denoted as *B. cereus* biovar *anthracis* in order to emphasize the biological significance of constituting a cause of anthrax-like disease [5,6]. *B. cereus* is an opportunistic human pathogen able to cause a variety of human diseases, among which the most prominent are the diarrhoeal or emetic syndromes. *B. cereus* infections also include severe cases following trauma to the eye, as well as a range of other opportunistic infections [7,8]. The general virulence factors of *B. cereus*, except for the NRPS-encoding genes responsible for emetic toxin synthesis in emetic strains, are frequently encoded on the bacterial chromosome [9,10]. *B. thuringiensis* is pathogenic to insect larvae, and other invertebrates such as nematodes and mites, and is therefore widely used commercially as a biological insecticide [11,12]. The bacterium produces specific insecticidal protein

Abbreviations: cgMLST, Core genome multilocus sequence typing.

* Corresponding authors. Department of Pharmacology and Pharmaceutical Biosciences, University of Oslo, Norway. (Nicolas J. Tourasse), (Ole Andreas Økstad)
E-mail addresses: nicolas.tourasse@inserm.fr (N.J. Tourasse), keith.jolley@biology.ox.ac.uk (K.A. Jolley), a.b.kolsto@farmasi.uio.no (A.-B. Kolstø), aloechen@farmasi.uio.no (O.A. Økstad).

https://doi.org/10.1016/j.resmic.2023.104050

toxins responsible for entomopathogenicity, and which are frequently plasmid-encoded, but in addition carries many of the same chromosomal virulence factors as *B. cereus* [13].

More recently, a total of at least 25 new species have been proposed for the *B. cereus* group [14], and are used by DNA sequence repositories such as GenBank and the EMBL databases. Several of these species have been defined on the basis of average nucleotide identity (ANI) values [15,16], in some cases with only one isolate known for each new species suggested, which contradicts formal rules for attribution of the species status to bacteria.

Various DNA sequence-based methods have been developed for resolving the phylogeny of *B. cereus* group isolates, including multilocus sequence typing (MLST; [1,17–21]) and amplified fragment length polymorphism (AFLP; [2,22–24]), in addition to the protein-based method multilocus enzyme electrophoresis (MLEE; [25,26]). While multilocus sequence typing (MLST) makes use of a limited set of typically six or seven core genes present in all members of a species or group of bacteria, for phylogenetic reconstruction, core genome MLST (cgMLST) reconstructs a phylogeny based on the full set of core genes available for the species or group. cgMLST is therefore in general a more powerful method compared to conventional MLST, and provides higher resolution and discriminatory power [27,28]. This is particularly useful for outbreak investigations and epidemiological surveillance and cgMLST is becoming a method of choice for isolate typing in public health laboratories [29,30]. At present, cgMLST schemes are being developed for a growing number of bacteria (or groups of bacteria; https://pubmlst.org, https://cgmlst.org) including many of the bacterial pathogens. Whole-genome (and even pan-genome) MLST (wgMLST) schemes incorporating all genes (core and accessory) have been designed for some species to provide even higher resolution if required. Another MLST variant based on 53 universally conserved ribosomal protein genes (rMLST) is also applied to all bacteria ([31]; https://pubmlst.org/rmlst). For the *B.* cereus group, five conventional seven-gene MLST schemes have been published, along with a pan-genome wgMLST scheme based on 30,363 loci (which is available through a commercial service only; https://www.applied-maths.com/applications/wgmlst), and the pan-bacterial rMLST scheme. A cgMLST scheme is available specifically for *B. anthracis* [32]. Here we present the first cgMLST scheme for the complete *B. cereus* group, including *B. cytotoxicus* which forms an outgroup [33,34], and the implementation of the scheme in the PubMLST resource [35]. This provides an unprecedented resolution to phylogenetic relationships in the group and includes the prediction of a new core genome for *B. cereus* group bacteria, based solely on high quality closed genomes and employing state-of-the-art methodology and approach. The resulting cgMLST scheme has then been applied to all available genomes in the group, resulting in a phylogeny including 2458 isolates for which whole genome sequence data is publicly available.

2. Materials and methods

2.1. cgMLST scheme creation

In a typical cgMLST analysis, a scheme is first developed using complete (or nearly complete) genomes and then applied to all other genomes of a given species (or group of species). The latest assemblies of 2458 *B. cereus* group genomes available at the time of study (October 2021) were downloaded from the NCBI GenBank FTP site. Among those, 173 were complete (i.e., ungapped, fully closed chromosome and plasmid sequences; genome list in Supplementary Table S1). The complete genome set included representatives of 12 of the recognized and proposed *B. cereus* group species, with one to five genomes per species, except for the main species *B. anthracis*, *B. cereus*, and *B. thuringiensis* which had 50 representatives. The cgMLST scheme was defined using the chewBBACA 2.8.5 pipeline [36]. The following steps were performed: "CreateSchema", to create a gene-by-gene scheme based on the set of complete genomes; "AlleleCall", to determine the allelic profiles based on the scheme, and "ExtractCgMLST", to define the set of loci that constitute the core genome. Briefly, in "CreateSchema" the algorithm clusters the coding sequences (CDS) of the complete genomes into unique loci by BLAST score ratio (BSR) analysis [37]. Then, "AlleleCall" consists in determining the alleles of each locus in all complete genomes based on sequence identity, BSR score, and gene size. Notably, this step also allows the detection and removal of possibly paralogous loci (when a CDS shows sequence similarity to multiple loci). At a given locus, each unique sequence is given an arbitrary allele number and for each isolate the combination of allele numbers at all loci defines an allelic profile. In the "ExtractCgMLST" operation, a set of core loci present in a predefined proportion of the genomes is extracted. All three analyses were run using default parameters (including a BSR cut-off of 0.6 and 20% gene size difference), except for the minimal CDS length in "CreateSchema" which was set to 90 bases (option "–l 90") and the genes were selected to be part of the core genome if they were present in a threshold proportion of 99% of the complete genomes in "ExtractCgMLST" ("–t 0.99"). In the extraction stage, paralogous loci detected by "AlleleCall" and two genome assemblies showing the highest numbers of missing loci (GCA_002243685.1 and GCA_000724585.1) were excluded (using options "–r" and "–g"), as recommended in chewBBACA. ChewBBACA identifies the CDS by means of the Prodigal gene prediction tool [38], which requires a training phase to learn the coding properties of the input organism. For the sake of comparison and consistency, Prodigal was trained on the genome sequence of the *B. anthracis* Ames Ancestor strain, which was used as a reference for creating the *B. anthracis* cgMLST scheme [32]. Prodigal 2.6.3 was run with the options "-p single" and "-c".

It has been shown that the *B. cytotoxicus* species has a significantly smaller genome, and thus gene content, than other members of the *B. cereus* group (~1 Mbp smaller; ~1000 genes less; [34]). As the scheme was designed for the whole *B. cereus* group, a set of 76 genes that were missing only in *B. cytotoxicus* were removed from the scheme (genes listed in Supplementary Table S2).

For comparison, we also analyzed the core genome using the most recent pan-genome software, Panaroo [39]. The 173 closed *B. cereus* group genomes were first annotated using Prokka 1.14.6 [40] (run with options "–addgenes –usegenus –genus Bacillus –kingdom Bacteria –gcode 11 –evalue 1e-09 –coverage 50 –mincontiglen 200") to provide consistent annotation files (in GFF3 format including gene sequences) for Panaroo (note that Prokka also makes use of Prodigal to predict the CDS). Using these annotations, Panaroo 1.2.10 was run in strict mode ("–clean-mode strict") with a core threshold of 99% ("-a core –core_threshold 0.99"), the MAFFT 7.407 alignment program ("–aligner mafft") [41] and all other default options. Because the locus identifiers generated by chewBBACA and Panaroo are different, the correspondence between the loci common to the two programs were determined by searching for sequence identity between the allele sequences of *B. anthracis* Ames Ancestor in chewBBACA and Panaroo using BLASTN+ 2.9.0 (run with parameters "-evalue 1.0e-10 -dust no -word_size 7"; [42]). As Prodigal is used for CDS prediction both in chewBBACA and Prokka (needed for Panaroo), the sequences should match exactly.

2.2. Application of the cgMLST scheme to B. cereus group genomes

To evaluate the cgMLST scheme, allele calling ("AlleleCall" operation in chewBBACA) for the loci in the scheme was performed on the full set of 2458 *B. cereus* group genomes available

(Supplementary file S3). Nine loci that were found to be duplicated in more than 5% of all genomes (classified as non-informative paralogous hits - NIPH or NIPHEM allele call in chewBBACA) were stripped from the scheme (Supplementary Table S3). In addition, the "TestGenomeQuality" utility of chewBBACA was used to check the completeness of the genome assemblies by counting the number of core loci that can be recovered in 99% of the genomes as genomes containing more and more missing loci are added to the analysis. The options were set to 12 iterations, a maximum of 1050 missing loci per genome, and a step of 10 missing loci ("-n 12 -t 1050 -s 10").

To make the cgMLST scheme publicly accessible, maintained, and updated over time, the scheme was included in the central PubMLST database (https://pubmlst.org/organisms/bacillus-cereus; [35]). To build the allelic profiles, the loci sequences for the reference *B. anthracis* Ames Ancestor strain annotated in NCBI RefSeq were used (assembly accession GCF_000008445.1). In chewBBACA CDS sequences are predicted using Prodigal and are given custom identifiers. The correspondence between the chewBBACA/Prodigal allele sequences and the RefSeq CDS was found by first aligning the chewBBACA/Prodigal alleles for the *B. anthracis* Ames Ancestor strain to the *B. anthracis* Ames Ancestor genome sequence using BLASTN+ 2.9.0 (with parameters "-evalue 1.0e-10 -dust no -word_size 7") to retrieve their genomic coordinates, and then by intersecting those coordinates with the coordinates of the CDS annotated in RefSeq using the "intersectBed" utility of the BEDTools 2.30.0 package (with options "-f 0.50 -r -loj"; [43]). As expected, there was a unique and complete or large overlap between the chewBBACA/Prodigal and GenBank loci for virtually every gene in the cgMLST scheme. Five loci that did not overlap with the GenBank annotation were deleted from the scheme (listed in Supplementary Table S4).

Starting from the loci sequences of the reference strain *B. anthracis* Ames Ancestor, homolog identification and allele calling for all loci in >2500 *B. cereus* group genomes were performed using the automated computational pipelines of pubMLST which are based on BLAST analysis [35].

2.3. Reconstruction of cgMLST phylogenetic tree

A phylogenetic tree was reconstructed using GrapeTree [44] and the matrix of allelic profiles for 2458 *B. cereus* group genomes generated by chewBBACA. The FastME V2 (Fast Minimum Evolution) Neighbor-Joining [45] pairwise distance method was employed (option "–method NJ" in GrapeTree). Missing loci were ignored specifically for each pairwise comparison of profiles, which is the default behavior in GrapeTree. The output tree (in Newick format) was then converted to phyloXML format (using the "phyloxml_converter" utility of the "forester" software libraries, run with option "-f = sn"; https://www.phylosoft.org/forester/) and subsequently imported into the iTOL web server [46] to annotate and color-code the tree figure according to the species and the seven major clusters defined by Guinebretiere and co-workers [2].

3. Results and discussion

3.1. A core-genome MLST scheme for the B. cereus group

In this study we created a core-genome MLST scheme for the whole *B. cereus* group. The scheme was built using the MLST-dedicated chewBBACA software [36] on the basis of 173 complete and closed *B. cereus* group genomes, and consists of 1568 genes that are conserved in 99% or more of those genomes (Supplementary Table S5; Supplementary file S1). The chewBBACA pipeline identifies homologous alleles and assigns allele numbers only to loci that satisfy a number of criteria, in particular a defined size range, sequence identity threshold to homologs, and that do not show duplication within loci or cross-homology between loci. Thus the 1568 selected core genes are those for which an unambiguous allele could be assigned in 99% or more of the closed genomes. Allele assignment in the available collection of 2458 *B. cereus* group genomes, which includes drafts of varying completeness and assembly quality, showed that 1463 of the core loci (93%) can be properly assigned in 99% of 2419 genomes with a high degree of completeness (that lack less than 10% of the loci), indicating that the scheme is overall valid for the *B. cereus* group (Fig. 1; Supplementary Fig. S1). It should be emphasized that for missing loci (here defined as loci for which an allele could not be unambiguously assigned in a given strain) we cannot rule out the possibility that they may not be truly missing from the isolates, but could be part of undetermined sequence gaps, especially for highly fragmented and incomplete genomes (e.g. consisting of several hundred contigs or more). Note that 17 of the 25 genes used in regular MLST schemes for the *B. cereus* group are retained in the core as defined here, and included in the cgMLST scheme, indicating that 8 of the genes previously selected for single regular MLST schemes are absent in more than 1% of the 173 closed genomes used to define the core genes.

cgMLST analyses are typically assembly-based (as is the case in pubMLST) and most assemblies are unfinished (i.e., not closed). Genome quality is judged with metrics that reflect contiguity (such as the N_{50} value or the number and length of contigs) or proper gene identification (e.g., the number of "good" cgMLST gene targets present). Genomes with high numbers of missing loci may introduce uncertainties in phylogenetic analyses and may be excluded or, if included, their placement should be taken with caution (for the current cgMLST scheme, genomes are flagged as "good" in pubMLST if they contain $\leq$5% missing targets). The possible effect of nucleotide errors in sequences is rarely considered during cgMLST analyses. However, the vast majority of bacterial genomes to date were sequenced using high coverage and low error rate short-read Illumina methodology. In the case of the *B. cereus* group, among the 2215 genomes for which information about sequencing technology and genome coverage was readily available (in the NCBI GenBank files) at the time of the present analysis, 91% are Illumina-derived and 95% have a sequencing depth of 20X or higher. Thus, the number of sequencing errors would be assumed to be small, and even if some gene sequences contain errors (leading to a wrong allele assignment) this should have little impact on the phylogenetic reconstructions, and certainly smaller impact compared to conventional MLST, given the large number of loci used in cgMLST. Genomes derived solely from long-read and/or low coverage data could potentially contain a higher number of nucleotide errors, causing uncertainties if many loci are affected. Assembly-free methods based on direct mapping of sequencing reads onto reference allele sequences have also been developed (e.g., SRST [47] and MOST [48]), which were shown to be more accurate and sensitive than assembly-based approaches for allele assignment, even for low coverage genomic data.

The cgMLST scheme presented here was designed to encompass all species in the *B. cereus* group. Therefore, when creating the scheme, we have purposely excluded a set of 76 additional genes that were present in 99% of the *B. cereus* group strains, but with the systematic exception of the *B. cytotoxicus* clade (Supplementary Table S2). *B. cytotoxicus* is recognized as a divergent member of the *B. cereus* group, and also has a significantly smaller (~1 Mbp smaller) genome compared to the other members of the group [34]. Although a single *B. cytotoxicus* isolate was included among the closed genomes used to identify the core genome, applying the cgMLST scheme to all *B. cereus* group genomes revealed that 70 out

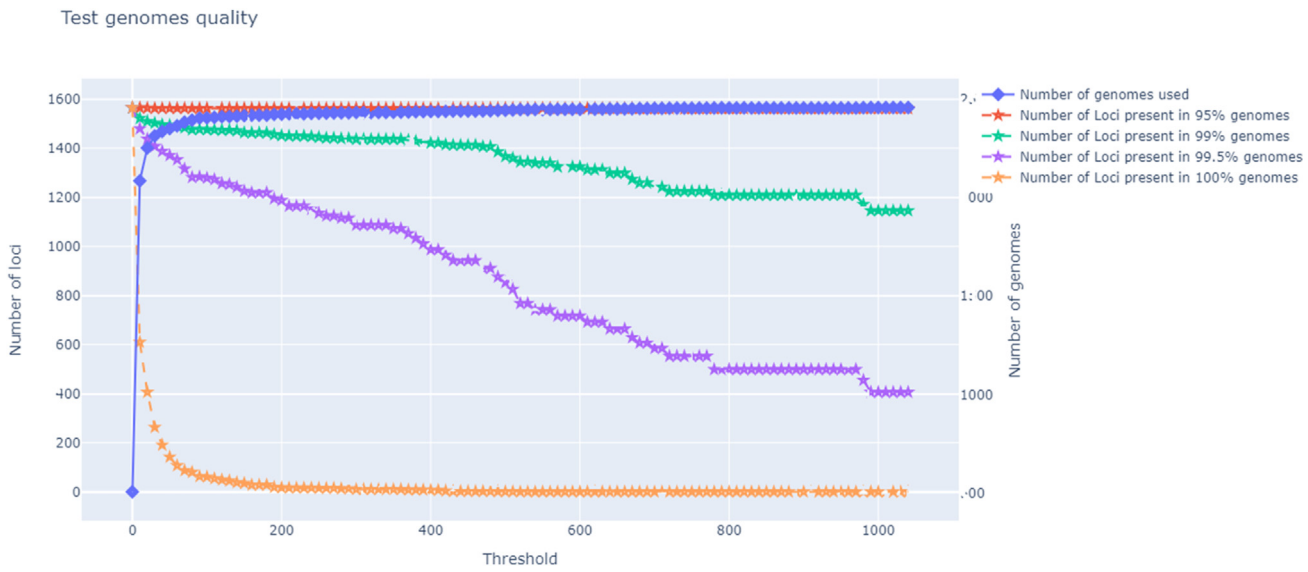

Fig. 1. Conservation of core loci as a function of genome completeness. Each point in the dashed lines gives the number of core loci in the cgMLST scheme that can be recovered in predefined proportions of *B. cereus* group genomes that are lacking an increasing threshold number of loci (x-axis threshold = max. number of missing loci). The continuous blue line gives the number of genomes for each threshold. The analysis was done using the "TestGenomeQuality" utility of chewBBACA [36] and shows that 1463 of the 1568 core loci included in the cgMLST scheme (93%) can be properly identified in 99% of 2419 good quality genomes (i.e. defined as genomes that lack less than 10% of the core loci, i.e., 160 loci). An interactive version of the figure is provided in Supplementary Data (Supplementary Fig. S1).

of the 76 genes were indeed lacking in all 14 *B. cytotoxicus* strains for which genomes are available. These genes thus appear to be specifically missing in the *B. cytotoxicus* species, and should not be considered as being part of the *B. cereus* group core genome. These genes included several transcriptional regulators, one two-component system, one HAMP domain-containing histidine kinase, and a range of transporter proteins (Supplementary Table S2). What impact the lack of these 70 proteins has on the biology of isolates belonging to the *B. cytotoxicus* clade is more difficult to interpret, as functional data is limited. Clearly however, the proportion of genes not found in this clade altogether, relative to the rest of the *B. cereus* group members, is biased towards the accessory gene pool rather than the core (76 genes specifically missing from the 1568 gene core (4.8%), while the ~1 Mb smaller genome found in *B. cytotoxicus* represents an approximate 20% reduction in genome size compared to other *B. cereus* group species).

In a thorough and comprehensive phylogenomic analysis of the *B. cereus* group, Bazinet inferred, using the Roary software [49] and a set of 114 closed genomes, that the core genome of the group was comprised of 598 genes [3], 416 of them being common to our cgMLST scheme. However, it was later shown by Tonkin-Hill and colleagues that Roary greatly underestimates the size of core genomes compared to more recent pan-genome analysis softwares [39]. The proposed explanation is that in Roary, clusters of homologous genes tend to be incorrectly split into multiple smaller clusters because of a too stringent pairwise identity threshold. Therefore, a core locus present in nearly all strains may not be identified as such because it is split into multiple loci present in smaller subsets of strains. In the course of this study we also tested Panaroo, described as the currently most accurate pan-genome tool [39]. When run on the set of 173 closed *B. cereus* group genomes, Panaroo output a core genome of 1905 genes present in at least 99% of the genomes (Supplementary file S2). However, 453 of the Panaroo core loci show sequence duplications in more than 1% of the closed genomes (i.e., 2 or more out of 173 genomes), and thus, if we apply the same criteria as in chewBBACA, no allele would be assigned for these loci and only the remaining 1452 Panaroo loci would be retained to create a cgMLST scheme. Out of these, 1182 are shared with our scheme. The core gene sets, as defined by ChewBBACA and Panaroo respectively, are therefore of overall largely comparable sizes, and overlapping by 75–81%. We have also compared our core gene set with the set obtained in the recent study of White et al. [50], in which the aim was to identify genes under different modes of evolutionary selection within the *B. cereus* group core and accessory gene pool. Here the authors used yet another pan-genome tool, PIRATE [51], and selected 1004 loci that were present in all (i.e., 100%) of 328 complete (or near complete) genomes, representing the "strict" core. Our cgMLST scheme includes 600 of these loci. It is however important to stress that the aim of our work was to develop a robust cgMLST scheme for use in typing and phylogenetic analysis of the *B. cereus* group, with no expectation of identifying a "true" core genome. Core gene identification is highly dependent on the set of genomes and bioinformatic tools used, and is very sensitive to the parameters and thresholds employed to select the genes. In this respect it is interesting to note that the remaining 404 core genes from the work of White and co-workers [50], the remaining 182 core genes from the analysis of Bazinet [3], and the remaining 763 core genes from the analysis done with Panaroo, and which are not overlapping with the core genes as defined here, are usually detected by chewBBACA just below our cutoff of presence in at least 99% of closed genomes. A few of the genes show duplications or wrong sizes in a large number of isolates, but most of the loci are not assigned by chewBBACA in only a few of the closed genomes (2–5 out of 173, i.e., a 97–99% window), and thus would be part of a "soft" core (defined as 95–99% presence frequency).

The new cgMLST scheme, employing the set of core genes as identified by chewBBACA, was then applied to all available *B. cereus* group genomes. The *B. cereus* group cgMLST scheme data generated here have been implemented and made available in the existing *B. cereus* section of the PubMLST website (https://pubmlst.org/organisms/bacillus-cereus; [35]), with the re-assignment of alleles. The scheme is available alongside the *B. anthracis* cgMLST scheme [32] and the widely-used conventional seven-gene MLST scheme developed by Priest and colleagues [17]. PubMLST provides a number of tools to browse, search, and retrieve allelic profiles and

sequences and isolate information, as well as tools to analyze and visualize phylogenetic relationships among isolates, and integration of the *B. cereus* cgMLST scheme in PubMLST will allow systematic maintenance and update of data, and ensure public availability of the scheme to the international community over time.

3.2. Gene content of the B. cereus group cgMLST scheme and overlap with the B. anthracis cgMLST scheme

From a comparison of the *B. cereus* group core genome, as defined in our analysis (1568 genes), to the core genome of *B. anthracis* (3803 genes), as defined from previous cgMLST analysis [32], 1225 out of the *B. cereus* group core genes defined here are also part of the *B. anthracis* core used for cgMLST, while unexpectedly 343 loci are not (Supplementary Table S6). Among these missing genes, we recognized that the *B. anthracis* scheme lacks a number of universal genes that are common in core bacterial genomes. These include e.g. specific translation elongation (Tu and Ts) and initiation (IF-1 and IF-3) factors and 26 ribosomal proteins, which are otherwise conserved in the full set of isolates belonging to the *B. cereus* group, as well as in other model bacterial species, and also 5 out of the 25 genes employed in the five conventional MLST schemes for the group (Supplementary Table S6). The reason for this discrepancy appears to be due to the fact that the *B. anthracis* cgMLST scheme was designed on the basis of 57 reference genomes, 31 of which were incomplete (i.e., not closed) [32]. In particular, a subset of ten genomes each lack between 50 and 80 of the *B. cereus* group core cgMLST loci.

3.3. Phylogenetic reconstruction of 2458 isolates from the B. cereus group with a complete genome sequence employing the devised cgMLST scheme

A phylogenetic tree was reconstructed using the GrapeTree software from pairwise comparisons between the allelic profiles of the 2458 *B. cereus* group isolates for which a genome sequence was available at the time of study (Fig. 2). The branches of the tree were colored according to the seven major clusters defined by Guinebretiere and colleagues (clusters II and III corresponding to clade 1, cluster IV corresponding to clade 2, and clusters I, V, VI, and VII corresponding to clade 3) [2,3]. The colors were set for isolates that were unambiguously assigned to a given cluster based on MLST, AFLP, and MLEE analyses, following previously published work [2,52]. As can be observed from the topology of the tree, isolates that were previously classified in the same clusters are included in monophyletic groups, demonstrating that the cgMLST-based phylogeny retains the main phylogenetic structure of the *B. cereus* group population (Fig. 2). As sequence divergence among bacterial isolates can come from individual point mutations or homologous recombination events (that can result in several mutations at once), in MLST, sequence variants are treated as different alleles (and assigned unique allele numbers) regardless of the number of nucleotide differences. Numerous tools exist to infer recombination in bacteria (reviewed in [53]). Standard MLST analyses have shown that the *B. cereus* group population is mainly clonal with limited recombination [1,17].

The cgMLST scheme here at its inception incorporates 2458 isolates, only about half of the number of isolates typed by the

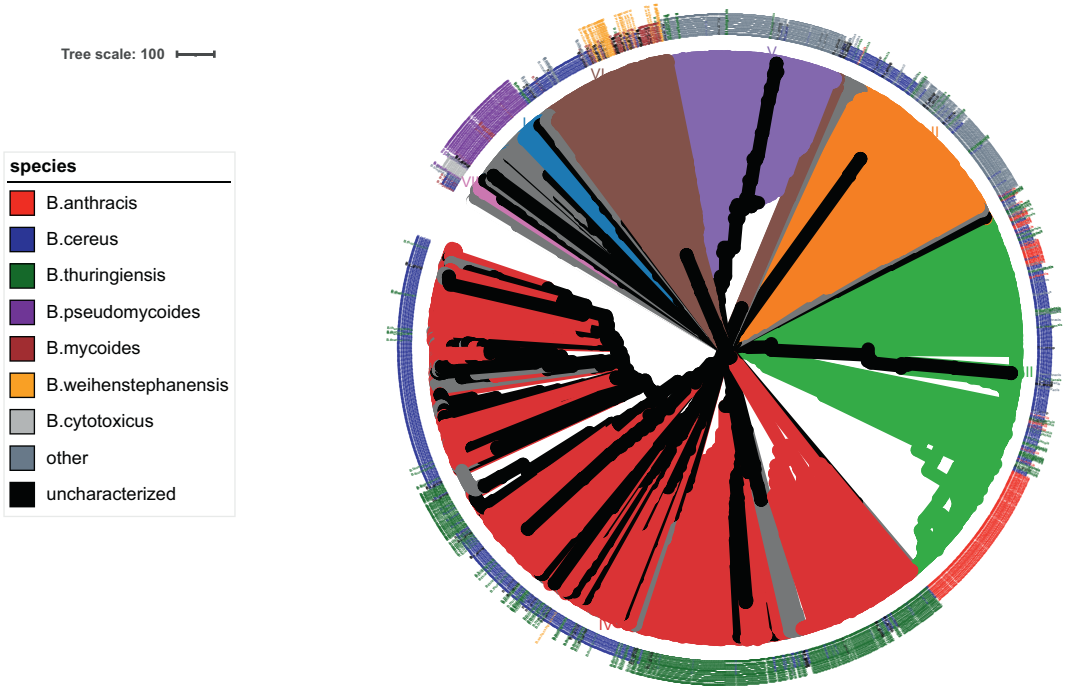

Fig. 2. Core-genome *B. cereus* group phylogenetic tree based on 2458 available genomes. The tree was reconstructed from a distance matrix between 2363 distinct cgMLST profiles using the FastME V2 Neighbor-Joining [45] method and the GrapeTree software [44]. Branches of the tree are color-coded according to the seven major clusters defined by Guinebretiere et al. [2], indicated by Roman numbers (I-VII). The colors were set for isolates that were previously assigned to a particular cluster based on MLST, AFLP, and MLEE analyses. Species are color-coded on the outer circle, as indicated. The tree was drawn in iTOL [46] and can be browsed interactively at https://itol.embl.de/shared/wYoEDoWlZcu1. The tree scale indicates the number of allelic differences.

conventional 6–7 gene MLST schemes (e.g. 5332 in PubMLST Oct 12, 2022; https://pubmlst.org/organisms/bacillus-cereus). With the continuous development in sequencing technologies and associated lower sequencing costs however, and thereby increasing numbers of isolates/strains for which whole genome sequence data is available, this situation will undoubtedly change proceeding from here. The development of a cgMLST scheme for the full *B. cereus* group is therefore timely, ensuring high-resolution strain typing of *B. cereus* group isolates in an open and fully public accessible online user interface for the future.

Supplementary data files

The supplementary files (Supplementary files S1, S2 and S3) were deposited on the Mendeley Data repository at https://data.mendeley.com/ (https://doi.org/10.17632/yd7n6xygvb.1).

Declaration of competing interest

The authors state that they do not have any conflict of interest.

Acknowledgements

We thank the University of Bordeaux for access to the Curta supercomputer of the "Mésocentre de Calcul Intensif Aquitain" (MCIA), and the "Institut Français de Bioinformatique" (IFB) for access to the IFB-core computer cluster.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.resmic.2023.104050.

References

[1] Helgason E, Tourasse NJ, Meisal R, Caugant DA, Kolstø AB. Multilocus sequence typing scheme for bacteria of the Bacillus cereus group. Appl Environ Microbiol 2004;70. https://doi.org/10.1128/AEM.70.1.191-201.
[2] Guinebretière MH, Thompson FL, Sorokin A, Normand P, Dawyndt P, Ehling-Schulz M, et al. Ecological diversification in the Bacillus cereus Group. Environ Microbiol 2008;10. https://doi.org/10.1111/j.1462-2920.2007.01495.x.
[3] Bazinet AL. Pan-genome and phylogeny of Bacillus cereus sensu lato. BMC Evol Biol 2017;17. https://doi.org/10.1186/s12862-017-1020-1.
[4] Head BM, Rubinstein E, Meyers AFA. Alternative pre-approved and novel therapies for the treatment of anthrax. BMC Infect Dis 2016;16. https://doi.org/10.1186/s12879-016-1951-y.
[5] Leendertz FH, Ellerbrok H, Boesch C, Couacy-Hymann E, Mätz-Rensing K, Hakenbeck R, et al. Anthrax kills wild chimpanzees in a tropical rainforest. Nature 2004;430:451–2. https://doi.org/10.1038/NATURE02722.
[6] Antonation KS, Grützmacher K, Dupke S, Mabon P, Zimmermann F, Lankester F, et al. Bacillus cereus biovar anthracis causing anthrax in sub-saharan Africa-chromosomal monophyly and broad geographic distribution. PLoS Negl Trop Dis 2016;10. https://doi.org/10.1371/JOURNAL.PNTD.0004923.
[7] Drobniewski FA. Bacillus cereus and related species. Clin Microbiol Rev 1993;6. https://doi.org/10.1128/cmr.6.4.324-338.
[8] Bottone EJ. Bacillus cereus, a volatile human pathogen. Clin Microbiol Rev 2010;23. https://doi.org/10.1128/CMR.00073-09.
[9] Ivanova N, Sorokin A, Anderson I, Galleron N, Candelon B, Kapatral V, et al. Genome sequence of Bacillus cereus and comparative analysis with Bacillus anthracis. Nature 2003;423. https://doi.org/10.1038/nature01582.
[10] Rasko DA, Ravel J, Økstad OA, Helgason E, Cer RZ, Jiang L, et al. The genome sequence of Bacillus cereus ATCC 10987 reveals metabolic adaptations and a large plasmid related to Bacillus anthracis pXO1. Nucleic Acids Res 2004;32. https://doi.org/10.1093/nar/gkh258.
[11] Crickmore N. Beyond the spore - past and future developments of Bacillus thuringiensis as a biopesticide. J Appl Microbiol 2006;101. https://doi.org/10.1111/j.1365-2672.2006.02936.x.
[12] Jouzani GS, Valijanian E, Sharafi R. Bacillus thuringiensis: a successful insecticide with new environmental features and tidings. Appl Microbiol Biotechnol 2017;101. https://doi.org/10.1007/s00253-017-8175-y.
[13] Ehling-Schulz M, Lereclus D, Koehler TM. The Bacillus cereus group: Bacillus species with pathogenic potential. Microbiol Spectr 2019;7. https://doi.org/10.1128/MICROBIOLSPEC.GPP3-0032-2018.
[14] Carroll LM, Cheng RA, Wiedmann M, Kovac J. Keeping up with the Bacillus cereus group: taxonomy through the genomics era and beyond. Crit Rev Food Sci Nutr 2021. https://doi.org/10.1080/10408398.2021.1916735.
[15] Liu B, Liu GH, Hu GP, Cetin S, Lin NQ, Tang JY, et al. 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 2014;105. https://doi.org/10.1007/s10482-013-0102-3.
[16] Liu Y, Du J, Lai Q, Zeng R, Ye D, Xu J, et al. Proposal of nine novel species of the bacillus cereus group. Int J Syst Evol Microbiol 2017;67. https://doi.org/10.1099/ijsem.0.001821.
[17] Priest FG, Barker M, Baillie LWJ, Holmes EC, Maiden MCJ. Population structure and evolution of the Bacillus cereus group. J Bacteriol 2004;186. https://doi.org/10.1128/JB.186.23.7959-7970.2004.
[18] Tourasse NJ, Helgason E, Økstad OA, Hegna IK, Kolstø AB. The Bacillus cereus group: novel aspects of population structure and genome dynamics. J Appl Microbiol 2006;101. https://doi.org/10.1111/j.1365-2672.2006.03087.x.
[19] Ko KS, Kim JW, Kim JM, Kim W, Chung SI, Kim IJ, et al. Population structure of the Bacillus cereus group as determined by sequence analysis of six housekeeping genes and the plcR gene. Infect Immun 2004;72. https://doi.org/10.1128/IAI.72.9.5253-5261.2004.
[20] Candelon B, Guilloux K, Ehrlich SD, Sorokin A. Two distinct types of rRNA operons in the Bacillus cereus group. Microbiology 2004;150. https://doi.org/10.1099/mic.0.26870-0.
[21] Sorokin A, Candelon B, Guilloux K, Galleron N, Wackerow-Kouzova N, Ehrlich SD, et al. Multiple-locus sequence typing analysis of Bacillus cereus and Bacillus thuringiensis reveals separate clustering and a distinct population structure of psychrotrophic strains. Appl Environ Microbiol 2006;72. https://doi.org/10.1128/AEM.72.2.1569-1578.
[22] Keim P, Kalif A, Schupp J, Hill K, Travis SE, Richmond K, et al. Molecular evolution and diversity in Bacillus anthracis as detected by amplified fragment length polymorphism markers. J Bacteriol 1997;179. https://doi.org/10.1128/jb.179.3.818-824.1997.
[23] Ticknor LO, KolstØ AB, Hill KK, Keim P, Laker MT, Tonks M, et al. Fluorescent amplified fragment length polymorphism analysis of Norwegian Bacillus cereus and Bacillus thuringiensis soil isolates. Appl Environ Microbiol 2001;67. https://doi.org/10.1128/AEM.67.10.4863-4873.2001.
[24] Hill KK, Ticknor LO, Okinaka RT, Asay M, Blair H, Bliss KA, et al. Fluorescent amplified fragment length polymorphism analysis of Bacillus anthracis, Bacillus cereus, and Bacillus thuringiensis isolates. Appl Environ Microbiol 2004;70. https://doi.org/10.1128/AEM.70.2.1068-1080.
[25] Helgason E, Caugant DA, Olsen I, Kolstø AB. Genetic structure of population of *Bacillus cereus* and *B. thuringiensis* isolates associated with periodontitis and other human infections. J Clin Microbiol 2000;38. https://doi.org/10.1128/jcm.38.4.1615-1622.2000.
[26] Helgason E, Caugant DA, Lecadet MM, Chen Y, Mahillon J, Lövgren A, et al. Genetic diversity of Bacillus cereus/B. thuringiensis isolates from natural sources. Curr Microbiol 1998;37. https://doi.org/10.1007/s002849900343.
[27] Kimura B. Will the emergence of core genome MLST end the role of in silico MLST? Food Microbiol 2018;75. https://doi.org/10.1016/j.fm.2017.09.003.
[28] Maiden MCJ, Van Rensburg MJJ, Bray JE, Earle SG, Ford SA, Jolley KA, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol 2013;11. https://doi.org/10.1038/nrmicro3093.
[29] Schürch AC, Arredondo-Alonso S, Willems RJL, Goering RV. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene-based approaches. Clin Microbiol Infect 2018;24:350–4. https://doi.org/10.1016/J.CMI.2017.12.016.
[30] Uelze L, Grützke J, Borowiak M, Hammerl JA, Juraschek K, Deneke C, et al. Typing methods based on whole genome sequencing data. One Heal Outlook 2020;2:3. https://doi.org/10.1186/S42522-020-0010-1.
[31] Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, et al. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiology 2012;158:1005–15. https://doi.org/10.1099/MIC.0.055459-0.
[32] Abdel-Glil MY, Chiaverini A, Garofolo G, Fasanella A, Parisi A, Harmsen D, et al. A whole-genome-based gene-by-gene typing system for standardized high-resolution strain typing of bacillus anthracis. J Clin Microbiol 2021;59. https://doi.org/10.1128/JCM.02889-20.
[33] Lapidus A, Goltsman E, Auger S, Galleron N, Ségurens B, Dossat C, et al. Extending the Bacillus cereus group genomics to putative food-borne pathogens of different toxicity. Chem Biol Interact 2008;171. https://doi.org/10.1016/j.cbi.2007.03.003.
[34] Guinebretière MH, Auger S, Galleron N, Contzen M, de Sarrau B, de Buyser ML, et al. Bacillus cytotoxicus sp. nov. is a novel thermotolerant species of the Bacillus cereus group occasionally associated with food poisoning. Int J Syst Evol Microbiol 2013;63. https://doi.org/10.1099/ijs.0.030627-0.
[35] Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications [version 1; referees: 2 approved]. Wellcome Open Res 2018;3. https://doi.org/10.12688/wellcomeopenres.14826.1.
[36] Silva M, Machado MP, Silva DN, Rossi M, Moran-Gilad J, Santos S, et al. chewBBACA: a complete suite for gene-by-gene schema creation and strain identification. Microb Genom 2018;4. https://doi.org/10.1099/mgen.0.000166.

[37] Rasko DA, Myers GSA, Ravel J. Visualization of comparative genomic analyses by BLAST score ratio. BMC Bioinf 2005;6. https://doi.org/10.1186/1471-2105-6-2.
[38] Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinf 2010;11. https://doi.org/10.1186/1471-2105-11-119.
[39] Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol 2020;21. https://doi.org/10.1186/s13059-020-02090-4.
[40] Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014;30. https://doi.org/10.1093/bioinformatics/btu153.
[41] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 2013;30: 772–80. https://doi.org/10.1093/molbev/mst010.
[42] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinf 2009;10:1–9. https://doi.org/10.1186/1471-2105-10-421.
[43] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841–2. https://doi.org/10.1093/bioinformatics/btq033.
[44] Zhou Z, Alikhan NF, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, et al. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. Genome Res 2018;28. https://doi.org/10.1101/gr.232397.117.
[45] Lefort V, Desper R, Gascuel O. FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. Mol Biol Evol 2015;32. https://doi.org/10.1093/molbev/msv150.
[46] Letunic I, Bork P. Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 2021;49. https://doi.org/10.1093/nar/gkab301.
[47] Inouye M, Dashnow H, Raven LA, Schultz MB, Pope BJ, Tomita T, et al. SRST2: rapid genomic surveillance for public health and hospital microbiology labs. Genome Med 2014;6. https://doi.org/10.1186/S13073-014-0090-6.
[48] Tewolde R, Dallman T, Schaefer U, Sheppard CL, Ashton P, Pichon B, et al. MOST: a modified MLST typing tool based on short read sequencing. PeerJ 2016;4. https://doi.org/10.7717/PEERJ.2308.
[49] Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 2015;31. https://doi.org/10.1093/bioinformatics/btv421.
[50] White H, Vos M, Sheppard SK, Pascoe B, Raymond B. Signatures of selection in core and accessory genomes indicate different ecological drivers of diversification among Bacillus cereus clades. Mol Ecol 2022;31:3584–97. https://doi.org/10.1111/mec.16490.
[51] Bayliss SC, Thorpe HA, Coyle NM, Sheppard SK, Feil EJ. PIRATE: a fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. Gigascience 2019;8. https://doi.org/10.1093/gigascience/giz119.
[52] Tourasse NJ, Okstad OA, Kolstø AB. HyperCAT: an extension of the SuperCAT database for global multi-scheme and multi-datatype phylogenetic analysis of the Bacillus cereus group population. Database 2010;2010. https://doi.org/10.1093/database/baq017.
[53] Shikov AE, Malovichko YV, Nizhnikov AA, Antonets KS. Current methods for recombination detection in bacteria. Int J Mol Sci 2022;23. https://doi.org/10.3390/IJMS23116257.