frontiers in Microbiology

ORIGINAL RESEARCH
published: 22 September 2020
doi: 10.3389/fmicb.2020.580691

No Assembly Required: Using BTyper3 to Assess the Congruency of a Proposed Taxonomic Framework for the *Bacillus cereus* Group With Historical Typing Methods

*Laura M. Carroll[1], Rachel A. Cheng[2] and Jasna Kovac[3]**

[1] Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany, [2] Department of Food Science, College of Agriculture and Life Sciences, Cornell University, Ithaca, NY, United States, [3] Department of Food Science, College of Agricultural Sciences, The Pennsylvania State University, University Park, PA, United States

OPEN ACCESS

Edited by:
Eric Altermann, AgResearch Ltd., New Zealand

Reviewed by:
Boris Alexander Vinatzer, Virginia Tech, United States
Justyna Malgorzata Drewnowska, University of Białystok, Poland
David W. Dyer, University of Oklahoma Health Sciences Center, United States

***Correspondence:**
Jasna Kovac
jzk303@psu.edu

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *06 July 2020*
Accepted: *25 August 2020*
Published: *22 September 2020*

Citation:
Carroll LM, Cheng RA and Kovac J (2020) No Assembly Required: Using BTyper3 to Assess the Congruency of a Proposed Taxonomic Framework for the Bacillus cereus Group With Historical Typing Methods. Front. Microbiol. 11:580691. doi: 10.3389/fmicb.2020.580691

The *Bacillus cereus* group, also known as *B. cereus sensu lato* (*s.l.*), is a species complex comprising numerous closely related lineages, which vary in their ability to cause illness in humans and animals. The classification of *B. cereus s.l.* isolates into species-level taxonomic units is essential for facilitating communication between and among microbiologists, clinicians, public health officials, and industry professionals, but is not always straightforward. A recently proposed genomospecies-subspecies-biovar taxonomic framework aims to provide a standardized nomenclature for this species complex but relies heavily on whole-genome sequencing (WGS). It thus is unclear whether popular, low-cost typing methods (e.g., single- and multi-locus sequence typing) remain congruent with the proposed taxonomy. Here, we characterize 2,231 *B. cereus s.l.* genomes using a combination of *in silico* (i) average-nucleotide identity (ANI)-based genomospecies assignment, (ii) ANI-based subspecies assignment, (iii) seven-gene multi-locus sequence typing (MLST), (iv) single-locus *panC* group assignment, (v) *rpoB* allelic typing, and (vi) virulence factor detection. We show that sequence types (STs) assigned using MLST can be used for genomospecies assignment, and we provide a comprehensive list of ST/genomospecies associations. For *panC* group assignment, we show that an adjusted, eight-group framework is largely, albeit not perfectly, congruent with the proposed eight-genomospecies taxonomy, as *panC* alone may not distinguish (i) *B. luti* from Group II *B. mosaicus* and (ii) *B. paramycoides* from Group VI *B. mycoides*. We additionally provide a list of loci that capture the topology of the whole-genome *B. cereus s.l.* phylogeny that may be used in future sequence typing efforts. For researchers with access to WGS, MLST, and/or *panC* data, we showcase how our recently released software, BTyper3 (https://github.com/lmc297/BTyper3), can be used to assign *B. cereus s.l.* isolates to taxonomic units within this proposed framework with little-to-no user intervention or domain-specific knowledge of *B. cereus s.l.* taxonomy. We additionally outline a novel

method for assigning *B. cereus s.l.* genomes to pseudo-gene flow units within proposed genomospecies. The results presented here highlight the backward-compatibility and accessibility of the recently proposed genomospecies-subspecies-biovar taxonomic framework and illustrate that WGS is not a necessity for microbiologists who want to use the proposed nomenclature effectively.

Keywords: *Bacillus anthracis*, *Bacillus cereus*, *Bacillus cereus* group, *Bacillus thuringiensis*, taxonomy, multi-locus sequence typing (MLST), whole-genome sequencing, single-locus sequence typing (SLST)

INTRODUCTION

The *Bacillus cereus* group, also known as *B. cereus sensu lato* (*s.l.*), is a species complex composed of numerous closely related, Gram-positive, spore-forming lineages with varying pathogenic potential (Rasko et al., 2005; Stenfors Arnesen et al., 2008; Messelhäußer and Ehling-Schulz, 2018; Ehling-Schulz et al., 2019). While some members of *B. cereus s.l.* have essential roles in agriculture and industry (e.g., as biocontrol agents) (Elshaghabee et al., 2017; Jouzani et al., 2017), others can cause illnesses with varying degrees of severity. Some members of the group, for example, are capable of causing severe forms of anthrax and anthrax-like illness that may result in death (Pilo and Frey, 2011, 2018; Moayeri et al., 2015). Other members of the group can cause foodborne illness that manifests in either an emetic form (i.e., intoxication characterized by vomiting symptoms and an incubation period of 0.5–6 h) or diarrheal form (i.e., toxicoinfection characterized by diarrheal symptoms and an incubation period of 8–16 h) (Stenfors Arnesen et al., 2008; Ehling-Schulz et al., 2015; Messelhäußer and Ehling-Schulz, 2018; Rouzeau-Szynalski et al., 2020).

Differentiating beneficial *B. cereus s.l.* strains from those that are capable of causing illness or death thus requires microbiologists, clinicians, public health officials, and industry professionals to communicate the potential risk associated with a given isolate. However, the lack of a "common language" for describing *B. cereus s.l.* isolates has hindered communication between and among scientists and other professionals and could potentially lead to dangerous mischaracterizations of an isolate's virulence potential. Anthrax-causing strains that possess phenotypic characteristics associated with "*B. cereus*" (e.g., motility, hemolysis on Sheep RBC) (Tallent et al., 2019), for example, have been referred to as "*B. anthracis*" (Leendertz et al., 2004), "*B. cereus*" (Hoffmaster et al., 2004; Avashia et al., 2007; Wilson et al., 2011), "*B. cereus* variety anthracis" (Klee et al., 2010), "*B. cereus* biovar anthracis" (Antonation et al., 2016; Marston et al., 2016), and "*B. cereus* biovar *anthracis*" or "*B. cereus* Biovar *anthracis*" (Brezillon et al., 2015; Antonation et al., 2016; Ehling-Schulz et al., 2019; Romero-Alvarez et al., 2020). Similarly, some *B. cereus s.l.* isolates that are closely related to emetic toxin (cereulide)-producing isolates are incapable of causing emetic intoxication themselves but can cause the diarrheal form of *B. cereus s.l.* illness (Ehling-Schulz et al., 2005; Jessberger et al., 2015; Riol et al., 2018; Carroll and Wiedmann, 2020). However, as there is no standardized name for these isolates, they have been referred to as "emetic-like *B. cereus*" (Ehling-Schulz et al., 2005), "*B. paranthracis*" (Liu et al., 2017; Bukharin et al., 2019), "*B. cereus,*" "Group III *B. cereus*" (i.e., assigned to Group III using the sequence of *panC* and the seven-phylogenetic group framework proposed by Guinebretiere et al., 2010), and "*B. cereus s.s.*" (although it should be noted that these strains do not fall within the genomospecies boundary of the *B. cereus s.s.* type strain and thus are not actually members of the *B. cereus s.s.* species) (Guinebretiere et al., 2010; Gdoura-Ben Amor et al., 2018; Glasset et al., 2018; Zhuang et al., 2019).

Recently, we proposed a standardized taxonomic nomenclature for *B. cereus s.l.* that is designed to minimize incongruencies and ambiguities within the *B. cereus s.l.* taxonomic space (Carroll et al., 2020b). The proposed taxonomy consists of: (i) a standardized set of eight genomospecies names (i.e., *B. pseudomycoides, B. paramycoides, B. mosaicus, B. cereus s.s., B. toyonensis*, *B. mycoides, B. cytotoxicus*, *B. luti*) that correspond to resolvable, non-overlapping genomospecies clusters obtained at a ≈92.5 average nucleotide identity (ANI) breakpoint; (ii) a formal collection of two subspecies names, which account for established lineages of medical importance within the *B. mosaicus* genomospecies (i.e., subspecies *anthracis*, which refers to the classic non-motile, non-hemolytic lineage known as "*B. anthracis,*" and subspecies *cereus*, which refers to lineages that encompass cereulide-producing isolates [i.e., "emetic *B. cereus*"] and the non-cereulide-producing isolates interspersed among them); and (iii) a standardized collection of biovar terms (i.e., Anthracis, Emeticus, Thuringiensis), which can be applied to any *B. cereus s.l.* isolate, regardless of genomospecies or taxonomic affiliation, to account for the heterogeneity of clinically and industrially important phenotypes (i.e., production of anthrax toxin, cereulide, and/or insecticidal crystal proteins, respectively) (Carroll et al., 2020b). However, this nomenclatural framework was developed using data derived from whole-genome sequencing (WGS) efforts, a technology that may not be accessible to all microbiologists or necessary for all microbiological studies. Hence, an assessment of congruency between WGS-based and single- or multi-locus sequencing-based genotyping and taxonomic assignment methods is needed. Here, we characterize 2,231 *B. cereus s.l.* genomes using a combination of *in silico* (i) ANI-based genomospecies assignment, (ii) ANI-based subspecies assignment, (iii) seven-gene multi-locus sequence typing (MLST), (iv) *panC* group assignment, (v) *rpoB* allelic typing, and (vi) virulence factor detection to show that popular, low-cost typing methods (e.g., single- and MLST) remain largely congruent with the proposed taxonomy. We additionally showcase how our recently released software, BTyper3 (Carroll et al., 2020b), can be used to assign *B. cereus s.l.* isolates to taxonomic units within this proposed

framework using WGS, MLST, and/or *panC* data. Further, we provide a list of loci that mirror the topology of the whole-genome *B. cereus s.l.* phylogeny, which may be used in future sequence typing efforts. Finally, we provide a novel method for assigning *B. cereus s.l.* isolates to pseudo-gene flow units using WGS data. The results presented here showcase that the proposed taxonomic framework for *B. cereus s.l.* is backward-compatible with historical *B. cereus s.l.* typing efforts and can be utilized effectively, regardless of whether WGS is used to characterize isolates or not.

MATERIALS AND METHODS

Acquisition of *Bacillus cereus s.l.* Genomes

All genomes submitted to the National Center for Biotechnology Information (NCBI) RefSeq (Pruitt et al., 2007) database as a published *B. cereus s.l.* species (Lechner et al., 1998; Guinebretiere et al., 2013; Jimenez et al., 2013; Miller et al., 2016; Liu et al., 2017; Carroll et al., 2020b) were downloaded (n = 2,231, accessed November 19, 2018; **Supplementary Table S1**). QUAST v. 5.0.2 (Gurevich et al., 2013) and CheckM v. 1.0.7 (Parks et al., 2015) were used to assess the quality of each genome, and BTyper3 v. 3.1.0 was used to assign each genome to a genomospecies, subspecies (if applicable), and biovar(s) (if applicable), using a recently proposed taxonomy (Carroll et al., 2020b). Genomes with (i) N50 > 100,000, (ii) CheckM completeness ≥ 97.5%, (iii) CheckM contamination ≤ 2.5%, and (iv) a genomospecies assignment that corresponded to a published *B. cereus s.l.* genomospecies were used in subsequent steps unless otherwise indicated (**Supplementary Table S1**). Genomes that did not meet these quality thresholds, as well as those which were assigned to an unknown or unpublished genomospecies (i.e., "Unknown *B. cereus* group Species 13–18" described previously) (Carroll et al., 2020b) or an effective or proposed *B. cereus s.l.* genomospecies (i.e., "*B. bingmayongensis,*" "*B. clarus,*" "*B. gaemokensis,*" or "*B. manliponensis,*" which were each assigned to a single genome), were excluded (Jung et al., 2010, 2011; Liu et al., 2014; Acevedo et al., 2019), yielding a set of 1,741 high-quality *B. cereus s.l.* genomes assigned to previously characterized, published *B. cereus s.l.* genomospecies. All subsequent analyses relied on one of two sets of genomes, as indicated: (i) the full set of 2,231 *B. cereus s.l.* RefSeq genomes, or (ii) the set of 1,741 high-quality genomes, with effective, proposed, unknown, and unpublished genomospecies removed. In some cases, the type strain genome of effective *B. cereus s.l.* species "*B. manliponensis*" was used to root a phylogeny, as it is the most distantly related member of the species complex (Jung et al., 2011; Carroll et al., 2020b).

Average Nucleotide Identity Calculations, Genomospecies Cluster Delineation, and Identification of Medoid Genomes

FastANI v. 1.0 (Jain et al., 2018) was used to calculate pairwise ANI values between all 1,741 high-quality *B. cereus s.l.* genomes (see section "Acquisition of *Bacillus cereus s.l.* Genomes" above). Genomospecies clusters and their respective medoid genomes were identified among all 1,741 genomes at all previously proposed ANI genomospecies thresholds for *B. cereus s.l.*, i.e., 92.5 (Carroll et al., 2020b), 94 (Jimenez et al., 2013), 95 (Guinebretiere et al., 2013; Miller et al., 2016), and 96 ANI (Liu et al., 2017), as described previously (Carroll et al., 2020b), using the bactaxR package (Carroll et al., 2020b) in R v. 3.6.1 (R Core Team, 2019) and the following dependencies: ape v. 5.3 (Paradis et al., 2004; Paradis and Schliep, 2019), cluster v. 2.1.0 (Maechler et al., 2019), dendextend v. 1.13.4 (Galili, 2015), dplyr v. 0.8.5 (Wickham et al., 2020), ggplot2 v. 3.3.0 (Wickham, 2016), ggtree v. 1.16.6 (Yu et al., 2017, 2018), igraph v. 1.2.5 (Csardi and Nepusz, 2006), phylobase v. 0.8.10 (R Hackathon, 2019), phytools v. 0.7-20 (Revell, 2012), readxl v. 1.3.1 (Wickham and Bryan, 2019), reshape2 v. 1.4.4 (Wickham, 2007), viridis v. 0.5.1 (Garnier, 2018).

FastANI was additionally used to calculate ANI values between each of the 2,231 genomes in the full set of *B. cereus s.l.* genomes and the type strain genomes of all 21 published and effective *B. cereus s.l.* species described prior to 2020 (**Supplementary Table S1**) so that the historical practice of assigning *B. cereus s.l.* genomes to species using type strain genomes could be assessed.

Construction of *B. cereus s.l.* Whole-Genome Phylogeny

To remove highly similar genomes and reduce the full set of 1,741 high-quality genomes to a smaller set of genomes that encompassed the diversity of *B. cereus s.l.* in its entirety, medoid genomes were identified among the set of 1,741 high-quality *B. cereus s.l.* genomes (see section "Acquisition of *Bacillus cereus s.l.* Genomes" above) at a 99 ANI threshold using the bactaxR package in R (see section "Average Nucleotide Identity Calculations, Genomospecies Cluster Delineation, and Identification of Medoid Genomes" above). Core single-nucleotide polymorphisms (SNPs) were identified among the resulting set of non-redundant genomes (n = 313; **Supplementary Table S1**) using kSNP3 v. 3.92 (Gardner and Hall, 2013; Gardner et al., 2015) and the optimal k-mer size reported by Kchooser (k = 19). IQ-TREE v. 1.5.4 (Nguyen et al., 2015) was used to construct a maximum likelihood phylogeny using the resulting core SNPs, the General Time-Reversible (Tavaré, 1986) nucleotide substitution model with a gamma rate-heterogeneity parameter (Yang, 1994) and ascertainment bias correction (Lewis, 2001) (i.e., the GTR + G + ASC nucleotide substitution model), and 1,000 replicates of the ultrafast bootstrap approximation (Minh et al., 2013; Hoang et al., 2018). The aforementioned core SNP detection and phylogeny construction steps were then repeated among the same set of 313 medoid genomes, with the addition of the "*B. manliponensis*" type strain genome (n = 314). The resulting phylogenies were annotated using the bactaxR package in R.

Construction of *panC*, *rpoB*, and Seven-Gene MLST Phylogenies

BTyper v. 2.3.2 (Carroll et al., 2017) was used to extract the nucleotide sequences of (i) *panC*, (ii) *rpoB*, and (iii) the seven

genes used in the PubMLST (Jolley and Maiden, 2010; Jolley et al., 2018) MLST scheme for *B. cereus* (i.e., *glp, gmk, ilv, pta, pur, pyc,* and *tpi*) from each of the 1,741 high-quality *B. cereus s.l.* genomes. MAFFT v. 7.453-with-extensions (Katoh et al., 2002; Katoh and Standley, 2013) was used to construct an alignment for each gene, and IQ-TREE was used to build a ML phylogeny from each resulting alignment, as well as an alignment constructed by concatenating the seven MLST genes, using the optimal nucleotide substitution model selected using ModelFinder (Kalyaanamoorthy et al., 2017) and 1,000 replicates of the ultrafast bootstrap approximation. The resulting phylogenies were annotated using the bactaxR package in R.

Construction of the Adjusted, Eight-Group *panC* Group Assignment Framework

Medoid genomes were identified among the full set of 1,741 high-quality *B. cereus s.l.* genomes at a 99 ANI threshold (n = 313; see section "Average Nucleotide Identity Calculations, Genomospecies Cluster Delineation, and Identification of Medoid Genomes" above). BTyper v. 2.3.3 was used to extract *panC* from each of the 313 *B. cereus s.l.* genomes, and MAFFT v. 7.453-with-extensions was used to construct an alignment. RhierBAPS v. 1.1.3 (Tonkin-Hill et al., 2018) was used to identify *panC* clusters within the alignment using two clustering levels; the nine top level (i.e., Level 1) clusters were used in subsequent steps, as they most closely mirrored the original seven *panC* groups (24 separate clusters were produced at Level 2). BTyper v. 2.3.3 was then used to extract *panC* from the full set of high-quality *B. cereus s.l.* genomes (n = 1,741; note that *panC* could not be extracted from all genomes), and the cd-hit-est command from CD-HIT v. 4.8.1 (Li and Godzik, 2006; Fu et al., 2012) was then used to cluster the resulting *panC* genes at a sequence identity threshold of 0.99. *panC* sequences that fell within the same CD-HIT cluster as a *panC* sequence from one or more of the 313 medoid genomes (n = 1,736) were assigned the RhierBAPS cluster of the medoid genome(s). MAFFT was used to construct an alignment of all 1,736 *panC* genes, and IQ-TREE v. 1.6.5 was used to construct a phylogeny using the resulting alignment as input, the optimal nucleotide substitution model selected using ModelFinder (i.e., the TVM + F + R4 model), and 1,000 replicates of the ultrafast bootstrap approximation.

The nine Level 1 RhierBAPS *panC* cluster assignments were then manually compared to *panC* groups assigned using BTyper v. 2.3.3 and the legacy seven-group framework. RhierBAPS *panC* groups were then re-named so that they most closely resembled the historical group assignments used by Guinebretiere et al. (2008, 2010) and BTyper v. 2.3.3 (Carroll et al., 2017).

Identification of Putative Loci for Future Single- and MLST Efforts

Prokka v. 1.12 (Seemann, 2014) was used to annotate each of the 313 *B. cereus s.l.* medoid genomes identified at 99 ANI (see section "Average Nucleotide Identity Calculations, Genomospecies Cluster Delineation, and Identification of Medoid Genomes" above), and the resulting protein sequences were divided randomly into 11 sets (ten sets containing 30 genomes, and one set containing 13 [the remainder] genomes) (Carroll et al., 2020b). OrthoFinder v. 2.3.3 (Emms and Kelly, 2015) was used to identify single-copy core genes present among all genomes in each set, and, subsequently, among all 313 genomes, using the iterative approach described previously (Carroll et al., 2020b). Nucleotide sequences of each of the 1,719 single-copy core genes present among all 313 genomes were aligned using MAFFT v. 7.453-with-extensions, and each resulting gene alignment was used as input for IQ-TREE v. 1.6.5. A maximum likelihood phylogeny was constructed for each gene using the GTR + G nucleotide substitution model and 1,000 replicates of the ultrafast bootstrap approximation.

The Kendall–Colijn (Kendall and Colijn, 2015, 2016; Jombart et al., 2017) test described by Katz et al. (2017) was used to assess the topological congruency between phylogenies constructed using each core gene and the "true" *B. cereus s.l.* whole-genome phylogeny (see section "Construction of *B. cereus s.l.* Whole-Genome Phylogeny" above). For each topological comparison, both phylogenies were rooted at the midpoint, and a lambda value of 0 (to give weight to tree topology rather than branch lengths) (Katz et al., 2017) and background distribution of 1,000 random trees were used. A phylogeny was considered to be more topologically similar to the "true" *B. cereus s.l.* whole-genome phylogeny than would be expected by chance if a significant P-value ($P < 0.05$) resulted after a Bonferroni correction was applied (Katz et al., 2017).

Metrics used for assessing the quality of putative typing loci included (i) length of the longest, uninterrupted/ungapped stretch of continuous sequence within the gene alignment, (ii) proportion of sites within the gene alignment that did not include gaps, (iii) proportion of the gene alignment that was covered by the longest uninterrupted/ungapped stretch of continuous sequence, and (iv) Bonferroni-corrected Kendall–Colijn P-value (**Supplementary Table S2**). Each individual gene was then detected within the full set of 1,741 high-quality *B. cereus s.l.* genomes (see section "Acquisition of *Bacillus cereus s.l.* Genomes" above) using nucleotide BLAST v. 2.9.0 (Camacho et al., 2009), as implemented in BTyper v. 2.3.3, by aligning the alleles of each single-copy core gene (n = 313) to each of the 1,741 genomes. A final set of candidate loci for single- and MLST was then identified (n = 255). A gene was included in the final set if: (i) $\geq$90% of the sites within the gene's alignment did not contain gap characters; (ii) the longest stretch of uninterrupted/ungapped continuous sequence within the gene's alignment covered $\geq$90% of the full length of the gene's alignment; (iii) the maximum likelihood phylogeny constructed using the gene as input was topologically similar to the "true" whole-genome phylogeny (i.e., Kendall–Colijn P-value < 0.05 after a Bonferroni correction); (iv) a single copy of the gene could be detected in all 1,741 high-quality *B. cereus s.l.* genomes, using minimum percent nucleotide identity and coverage thresholds of 90% each and a maximum E-value threshold of 1E-5 (**Supplementary Table S2**).

Functional Annotation of Putative Loci for Future Single- and MLST Efforts

Amino acid sequences of the resulting 255 candidate loci (see section "Identification of Putative Loci for Future Single- and

MLST Efforts"; **Supplementary Table S2**) were functionally annotated using eggNOG mapper v. 2.0 (Huerta-Cepas et al., 2017, 2019). The resulting Clusters of Orthologous Groups (COG) functional categories were visualized in R using the igraph v. 1.2.5 package (Csardi and Nepusz, 2006). The GOGO Webserver[1] was used to calculate pairwise semantic/functional similarities between genes based on their assigned Gene Ontology (GO) terms and to cluster genes based on their GO term similarities (Zhao and Wang, 2018). For each of the three GO directed acyclic graphs (i.e., Biological Process Ontology, Cellular Component Ontology, and Molecular Function Ontology) (Ashburner et al., 2000; The Gene Ontology Consortium, 2018), an $n \times n$ matrix of pairwise similarities produced by GOGO was converted into a dissimilarity matrix by subtracting all values from an $n \times n$ matrix containing 1.0s. Non-metric multidimensional scaling (NMDS) (Kruskal, 1964) was performed using the resulting dissimilarity matrix, the metaMDS function in the vegan (Oksanen et al., 2019) package in R, two dimensions ($k = 2$), and a maximum of 10,000 random starts. Convergent solutions were reached in under 100 random starts for the biological process and cellular component dissimilarity matrices and in under 1,400 random starts for the molecular function dissimilarity matrix. The results from each NMDS run were plotted in R using ggplot2.

Identification of Microbial Gene Flow Units Using Recent Gene Flow and Implementation of the Pseudo-Gene Flow Unit Assignment Method in BTyper3 v. 3.1.0

The "PopCOGenT" module available in PopCOGenT (downloaded October 5, 2019) (Arevalo et al., 2019) was used to identify gene flow units (i.e., "main clusters" reported by PopCOGenT) among the 313 *B. cereus s.l.* medoid genomes identified at 99 ANI (**Figure 1A**; see section "Average Nucleotide Identity Calculations, Genomospecies Cluster Delineation, and Identification of Medoid Genomes" above), using the following dependencies: Mugsy v. v1r2.3 (Angiuoli and Salzberg, 2011) and Infomap v. 0.2.0 (Rosvall et al., 2009).

Pairwise ANI values were then calculated between genomes within each of the 33 PopCOGenT gene flow units using FastANI v. 1.0, and bactaxR was used to identify the medoid genome for each PopCOGenT gene flow unit based on the resulting pairwise ANI values (**Figure 1A**). The minimum ANI value shared between the PopCOGenT gene flow unit medoid genome and all other genomes assigned to the same gene flow unit using PopCOGenT was treated as the observed ANI boundary for the gene flow unit; the observed ANI boundary formed by a PopCOGenT gene flow unit medoid genome forms what we refer to here as a pseudo-gene flow unit (**Figure 1A**).

The 33 resulting medoid genomes for each of the 33 pseudo-gene flow units, as well as the genomes of effective and proposed *B. cereus s.l.* species, were then used to create a rapid pseudo-gene flow unit typing scheme in BTyper3 v. 3.1.0 (**Figure 1**). For this approach, ANI values are calculated between a user's query genome and the set of 33 pseudo-gene flow unit medoid genomes using FastANI (**Figures 1B**, **2**). The closest-matching medoid genome and its ANI value relative to the query are identified; additionally, the previously observed ANI boundaries for the medoid genome's respective pseudo-gene flow unit are reported (**Figures 1B**, **2**). It is important to note that this pseudo-gene flow unit assignment method measures genomic similarity via ANI, which is fundamentally and conceptually very different from the methods that PopCOgenT employs. The ANI-based pseudo-gene flow unit assignment method described here does not query recent gene flow, nor does it use PopCOgenT or the methods that it employs directly. Thus, it cannot directly assign a genome to a PopCOgenT gene flow unit, and results should not be interpreted as a true measurement of gene flow. However, this approach allows researchers to rapidly identify the closest medoid genome of previously delineated true gene flow units (**Figure 1A**), based on a metric of genomic similarity, which provides insight into the phylogenomic placement of a query genome within a larger *B. cereus s.l.* genomospecies.

Implementation of Virulence Factor Detection in BTyper3 v. 3.1.0

Versions of BTyper3 prior to v. 3.1.0 (Carroll et al., 2020b), as well as the original BTyper (i.e., BTyper v. 2.3.3 and earlier) (Carroll et al., 2017) detected virulence factors using translated nucleotide BLAST (Camacho et al., 2009) and minimum amino acid identity and coverage thresholds of 50% and 70%, respectively, as these values had been shown to correlate with PCR-based detection of virulence factors (Kovac et al., 2016). However, these thresholds were selected using a limited number of *B. cereus s.l.* isolates with limited genomic diversity and can potentially lead to the detection of remote homologs that do not correlate with a virulence phenotype (i.e., false positive hits). For example, some *B. cereus s.l.* isolates possess a gene that shares a low degree of homology with *cesC*, but still meet these virulence factor detection thresholds (see **Figure 5** of Carroll et al., 2017) (Carroll et al., 2017). Users with limited knowledge of *B. cereus s.l.* virulence factors, or those who do not know how to interpret BLAST identity and coverage thresholds, may infer that these isolates have a potential to produce cereulide, when they actually do not (e.g., a distant *cesC* homolog is detected in *B. mycoides* type strain DSM 2048 at these thresholds, but *cesABD* are not detected, and the strain does not produce cereulide) (Ulrich et al., 2019). A similar phenomenon is observed with some members of the "*B. cereus*" exo-polysaccharide capsule (Bps)-encoding genes (e.g., *bpsEF*) (Carroll et al., 2019).

To provide updated boundaries for virulence factor detection based on a larger set of genomes that span *B. cereus s.l.*, BTyper3 v. 3.1.0 was used to identify virulence factors in the complete set of 1,741 high-quality genomes (see section "Acquisition of *Bacillus cereus s.l.* Genomes" above), using a maximum BLAST E-value threshold of 1E-5 (Carroll et al., 2017, 2020b), but with minimum amino acid identity and coverage thresholds of 0% each (**Supplementary Table S3**). Plots of virulence factors detected within all genomes at

[1] http://dna.cs.miami.edu/GOGO/; accessed May 30, 2020.

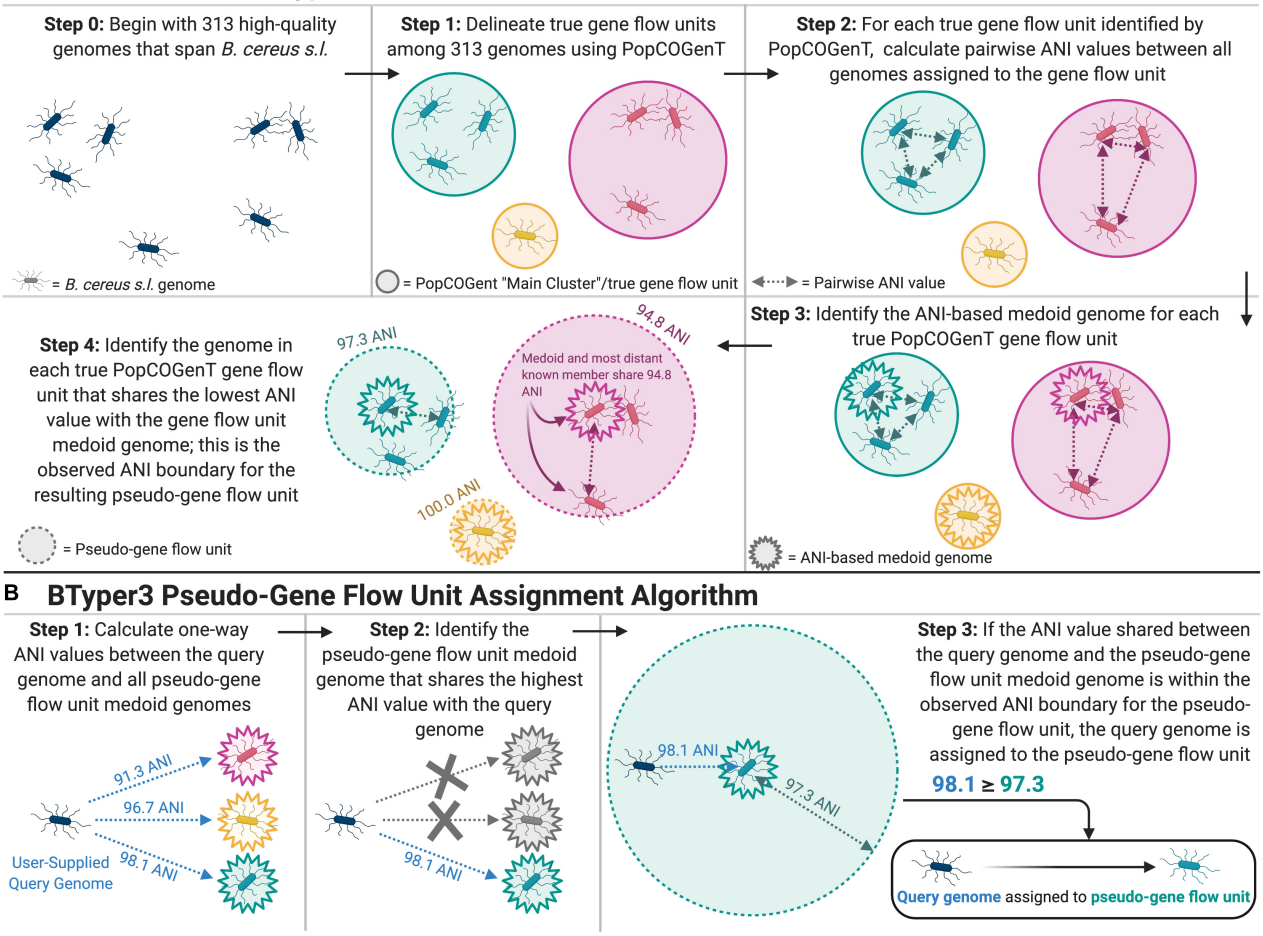

FIGURE 1 | (A) Graphical depiction of the methods used to construct *B. cereus s.l.* pseudo-gene flow units used by BTyper3. The 313 high-quality *B. cereus s.l.* genomes (Step 0) were medoid genomes identified among a set of 1,741 high-quality *B. cereus s.l.* genomes at a 99 average nucleotide identity (ANI) threshold using the bactaxR package in R. This step was performed to remove highly similar genomes and reduce the full set of 1,741 high-quality genomes to a smaller set of genomes that encompassed the diversity of *B. cereus s.l.* in its entirety. Gene flow units delineated using PopCOGenT (Step 1) were the "main clusters" reported by the PopCOGenT module. **(B)** Graphical depiction of the pseudo-gene flow unit assignment algorithm implemented in BTyper3. Pseudo-gene flow unit medoid genomes (Steps 1–3) are the output of the steps outlined in **(A)**. If a user-supplied query genome does not fall within the observed ANI boundary of the most similar pseudo-gene flow unit medoid genome (Step 3), the second-through-fifth most similar pseudo-gene flow unit medoid genomes are queried. All ANI values were calculated using FastANI. The figure was created with BioRender (https://biorender.com/).

various amino acid identity and coverage thresholds were constructed using ggplot2 in R (**Figure 3** and **Supplementary Table S3**). Based on these plots, amino acid identity and coverage thresholds of 70% and 80%, respectively, were implemented as the default thresholds for virulence factor detection in BTyper3 v. 3.1.0 (**Figure 3**). Additionally, to reduce the risk of users mis-interpreting spurious hits that do not correlate with a virulence phenotype, BTyper3 v. 3.1.0 reports virulence factors at an operon/cluster level; for example, if only *cesC* is detected in a genome, BTyper3 reports that only one of four cereulide synthetase-encoding genes were detected (**Figure 2**). Similarly, some *B. cereus s.l.* isolates possess genes that share a high degree of homology with Bps-encoding genes (e.g., >90% identity and coverage); to avoid users mis-interpreting that this isolate may produce a Bps capsule, BTyper3 reports the fraction of *bps* hits out of nine *bps* genes (**Figure 2**).

Implementation of Seven-Gene MLST in BTyper3 v. 3.1.0

The PubMLST seven-gene MLST scheme for *B. cereus s.l.* implemented in the original version of BTyper (i.e., BTyper v. 2.3.3 and earlier) was implemented in BTyper3 v. 3.1.0 as described previously (Carroll et al., 2017). The option to download the latest version of the *B. cereus s.l.* MLST database from PubMLST was also included in BTyper3 v. 3.1.0. Additionally, the clonal complex associated with each sequence type listed in PubMLST (if available), as well as the number of alleles that matched an allele in the PubMLST database with 100%

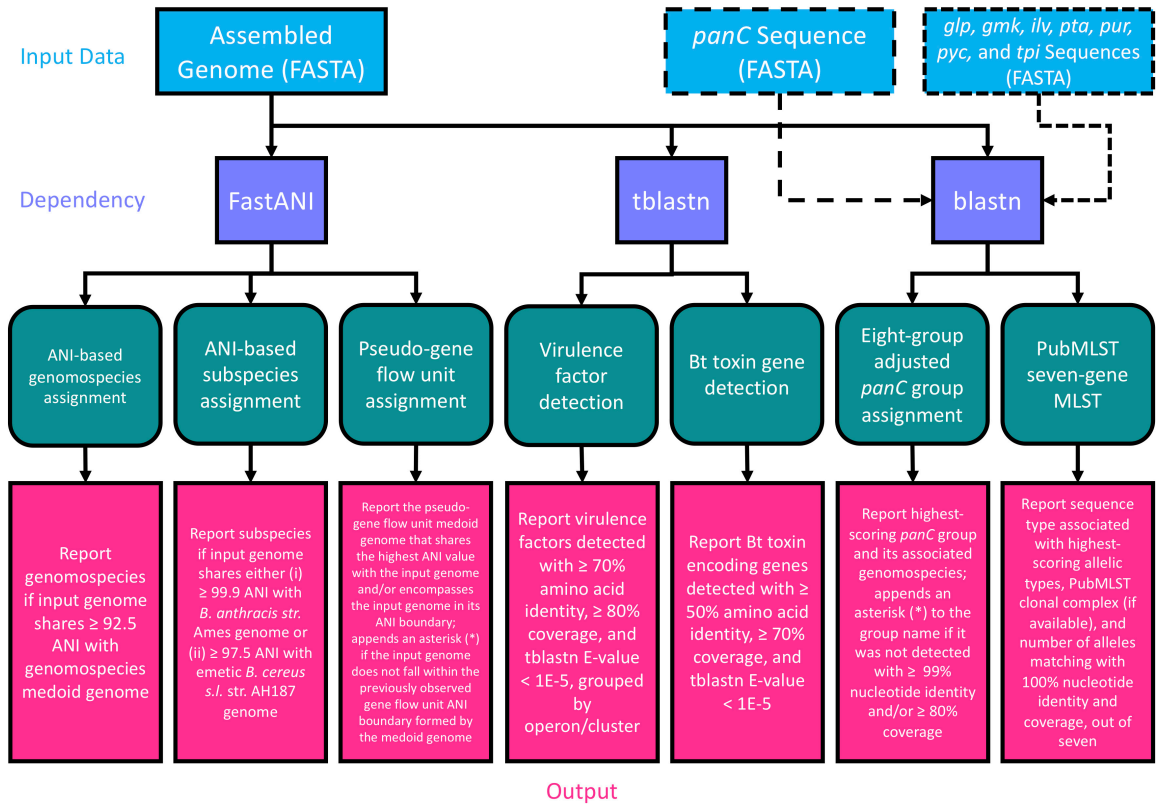

FIGURE 2 | Flow chart describing the workflow implemented in BTyper3 v. 3.1.0. Input data in FASTA format (blue boxes) can consist of any of the following: (i) a whole genome (complete or draft; can be used with all/all combinations of workflow steps), (ii) a *panC* sequence (can be used with the eight-group adjusted *panC* group assignment workflow only), or (iii) sequences of the seven loci used in PubMLST's seven-gene multi-locus sequence typing (MLST) scheme for *B. cereus s.l.* (sequences can be in multi-FASTA format, or concatenated into a single sequence; can be used with the PubMLST seven-gene MLST workflow only). Purple boxes represent software dependencies required for each type of input data, while green boxes represent the various analyses that can be conducted in BTyper3 v. 3.1.0. Pink boxes denote the output that BTyper3 v. 3.1.0 reports for each analysis.

identity and coverage out of seven, is reported in the BTyper3 final report (**Figure 2**).

Implementation of *panC* Group Assignment in BTyper3 v. 3.1.0

The updated eight-group *panC* group assignment framework developed here (see section "Construction of the Adjusted, Eight-Group *panC* Group Assignment Framework" above) was used to construct a typing method in BTyper3 v. 3.1.0 (**Figures 2**, **4**). Briefly, BTyper3 v. 3.1.0 assigns a genome to a *panC* group using a database of 64 representative *panC* sequences from the 1,736 *B. cereus s.l. panC* sequences clustered at a 99% identity threshold described above. *panC* sequences of effective and proposed *B. cereus s.l.* species are also included in the database but are assigned a species name (e.g., "Group_manliponensis") rather than a number (i.e., Group_I to Group_VIII). Nucleotide BLAST is used to align a query genome to the *panC* database, and the *panC* group producing the highest BLAST bit score is reported. Species associated with each *panC* group within the eight-group framework are also reported: (i) Group I (*B. pseudomycoides*), (ii) Group II (*B. mosaicus*/*B. luti*); (iii) Group III (*B. mosaicus*); (iv) Group IV (*B. cereus s.s.*); (v) Group V (*B. toyonensis*); (vi) Group VI (*B. mycoides*/*B. paramycoides*); (vii) Group VII (*B. cytotoxicus*); (viii) Group VIII (*B. mycoides*; **Figure 2**). If a query genome does not share ≥99% nucleotide identity and/or ≥80% coverage with one or more *panC* alleles in the database, the closest-matching *panC* group is reported with an asterisk (*).

BTyper3 Code Availability

BTyper3, its source code, and its associated databases are free and publicly available at https://github.com/lmc297/BTyper3.

RESULTS

Genomospecies Defined Using Historical ANI-Based Genomospecies Thresholds and Species Type Strains Are Each Integrated Into One of Eight Proposed *B. cereus s.l.* Genomospecies

Genomospecies assigned using higher, historical species cutoffs (i.e., 94, 95, and 96 ANI) and the type strain genomes of the 18 published *B. cereus s.l.* species described prior to 2020 were

TABLE 1 | Proposed genomospecies-level taxonomy for *B. cereus s.l.* isolates[a].

Proposed genomospecies name	Legacy *panC* group (I–VII)[b]	Adjusted *panC* group (I–VIII)[c]	Whole-genome sequencing (WGS)[d]
B. pseudomycoides	Group I	Group I	Shares $\gtrapprox$ 92.5 ANI with *B. pseudomycoides* str. DSM 12442^T (GCF_000161455.1)
B. mosaicus	Groups II/III	Groups II/III	Shares $\gtrapprox$ 92.5 ANI with *B. albus* str. N35-10-2^T (GCF_001884185.1), *B. anthracis* str. Ames (GCF_000007845.1), *B. mobilis* str. 0711P9-1^T (GCF_001884045.1), *B. pacificus* str. EB422^T (GCF_001884025.1), *B. paranthracis* str. MN5^T (GCF_001883995.1), *B. tropicus* str. N24^T (GCF_001884035.1), and/or *B. wiedmannii* str. FSL W8-0169^T (GCF_001583695.1)
B. cereus s.s.	Group IV	Group IV	Shares $\gtrapprox$ 92.5 ANI with *B. cereus s.s.* str. ATCC 14579^T (GCF_000007825.1) and/or *B. thuringiensis*. serovar berliner str. ATCC 10792 (GCF_000161615.1)
B. toyonensis	Group V	Group V	Shares $\gtrapprox$ 92.5 ANI with *B. toyonensis* str. BCT-7112^T (GCF_000496285.1)
B. mycoides	Groups II/III/VI	Groups VI/VIII	Shares $\gtrapprox$ 92.5 ANI with *B. mycoides* str. DSM 2048^T (GCF_000003925.1), *B. nitratireducens* str. 4049^T (GCF_001884135.1), *B. proteolyticus* str. TD42^T (GCF_001884065.1), *B. weihenstephanensis* str. WSBC 10204^T (GCF_000775975.1)
B. cytotoxicus	Group VII	Group VII	Shares $\gtrapprox$ 92.5 ANI with *B. cytotoxicus* str. NVH 391-98^T (GCF_000017425.1)
B. paramycoides	Group VI	Group VI	Shares $\gtrapprox$ 92.5 ANI with *B. paramycoides* str. NH24A2^T (GCF_001884235.1)
B. luti	Groups III/V/VI	Group II	Shares $\gtrapprox$ 92.5 ANI with *B. luti* str. TD41^T (GCF_001884105.1)

[a]*See* ***Supplementary Tables S5, S7*** *for multi-locus sequence typing (MLST) sequence types (STs) and rpoB allelic types (ATs) associated with each proposed genomospecies, respectively.* [b]*panC group assignment using the original BTyper (i.e., BTyper v. 2.3.3) and the legacy seven-group framework described by Guinebretiere et al. (2010); note that group assignments here, particularly for Groups II, III, and VI, may differ from those produced using the web-tool published by Guinebretiere et al. (2010), as the two methods rely on different panC databases.* [c]*panC group assignment using the adjusted eight-group panC framework described here.* [d]*Average nucleotide identity (ANI)-based comparisons to the type strain genomes of published species are described here, as B. cereus s.l. genomospecies classification prior to 2020 has relied on this practice/it is likely more meaningful to most B. cereus s.l. researchers. However, in practice, any genome of known genomospecies can be used for genomospecies assignment; see* ***Supplementary Table S1*** *for a complete list of genomospecies assignments for all B. cereus s.l. genomes (n = 2,231). Additionally, see* ***Supplementary Table S7*** *of Carroll et al. (2020b) for a list of medoid genomes for the above genomospecies.*

safely integrated into proposed genomospecies delineated at 92.5 ANI without polyphyly (**Supplementary Table S4**). Five of the eight genomospecies (i.e., *B. pseudomycoides, B. paramycoides, B. toyonensis, B. cytotoxicus,* and *B. luti*) encompassed all genomes assigned to the respective species using its type strain (**Table 1**), regardless of whether a 94, 95, or 96 ANI threshold was used. The remaining three genomospecies (i.e., *B. mosaicus, B. cereus s.s.,* and *B. mycoides*) simply integrated multiple species assigned using historical ANI-based genomospecies thresholds into a single genomospecies (**Table 1**). Regardless of whether a threshold of 94, 95, or 96 ANI was used, all genomes assigned to any of *B. albus, anthracis, mobilis, pacificus, paranthracis, tropicus,* and *wiedmannii* using species type strain genomes belonged to *B. mosaicus* (**Table 1** and **Supplementary Table S4**). Likewise, all genomes assigned to any of *B. mycoides, nitratireducens, proteolyticus,* and *weihenstephanensis* using species type strain genomes and genomospecies thresholds of 94–96 ANI were assigned to the *B. mycoides* genomospecies cluster (**Table 1** and **Supplementary Table S4**). Additionally, all genomes that shared 94–96 ANI with the *B. cereus s.s.* str. ATCC 14579 and/or *B. thuringiensis* serovar berliner str. ATCC 10792 type strain genomes belonged to the *B. cereus s.s.* genomospecies cluster (**Table 1** and **Supplementary Table S4**). However, it should be noted that the "*B. cereus*" and "*B. thuringiensis*" species as historically defined are polyphyletic, and other strains often referred to as "*B. cereus*" or "*B. thuringiensis*" belong to other genomospecies clusters; emetic reference strain "*B. cereus*" str. AH187, for example, belongs to *B. mosaicus* and not *B. cereus s.s.* (Carroll et al., 2019, 2020b).

STs Assigned Using Seven-Gene MLST Can Be Used for *B. cereus s.l.* Genomospecies Assignment

All STs assigned using BTyper3 and PubMLST's seven-gene MLST scheme *for B. cereus s.l.* (Jolley and Maiden, 2010; Jolley et al., 2018) were contained within a single proposed *B. cereus s.l.* genomospecies, and no STs were split across multiple genomospecies (**Supplementary Tables S1, S5**). As such, a comprehensive list of ST/genomospecies associations for all NCBI RefSeq *B. cereus s.l.* genomes is available (n = 2,231; RefSeq accessed November 19, 2018, PubMLST *B. cereus* database accessed April 26, 2020; **Supplementary Tables S1, S5**). However, it is essential to note that the *B. cereus s.l.* phylogeny constructed using the sequences of these seven alleles alone (i.e., the MLST phylogeny) did not mirror the WGS-based

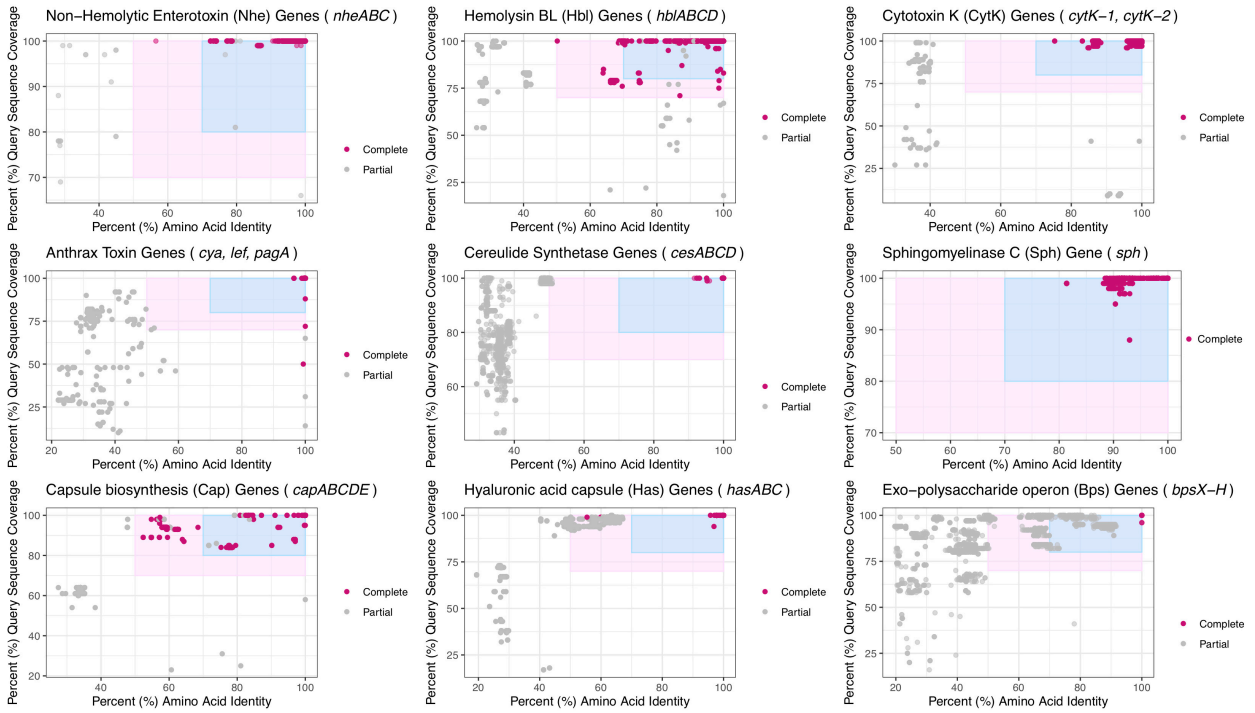

FIGURE 3 | Virulence factors detected in 1,741 high-quality *B. cereus s.l.* genomes at various minimum percent amino acid identity (*X*-axes) and query sequence coverage (*Y*-axes) thresholds. Each subplot denotes a virulence factor composed of one or more genes listed in the subplot title. Points represent the individual genes listed in the subplot title. The light pink rectangle denotes amino acid identity and coverage values at which the original BTyper (BTyper v. 2.3.3 and previous versions) would report a gene as "present" (i.e., 50 and 70% amino acid identity and coverage thresholds, respectively). The blue rectangle denotes the updated virulence factor cutoffs used by BTyper3 v. 3.1.0 (i.e., 70 and 80% amino acid identity and coverage thresholds, respectively). Points shaded in dark pink (i.e., "Complete") were (i) detected within a *B. cereus s.l.* genome at the default minimum amino acid identity and coverage thresholds used by the original BTyper (i.e., BTyper v. 2.3.3, at 50 and 70%, respectively), and (ii) were part of a "complete" virulence factor, as listed in the subplot title (i.e., all other genes comprising the virulence factor were detected in the genome at the 50 and 70% minimum amino acid identity and coverage thresholds used by BTyper v. 2.3.3, respectively). Points colored in gray (i.e., "Partial") denote genes that were not detected at the 50 and 70% minimum amino acid identity and coverage thresholds used by BTyper v. 2.3.3 and/or were part of a virulence factor that was not present in its entirety in the respective genome at 50 and 70% amino acid identity and coverage, respectively. All genes were detected using BTyper3 v. 3.1.0 with a minimum E-value threshold of 1E-5.

B. cereus s.l. phylogeny perfectly. Regardless of the ANI threshold used (i.e., 92.5, 94, 95, or 96 ANI), the *B. cereus s.l.* MLST phylogeny yielded polyphyletic genomospecies clusters (**Figure 5** and **Supplementary Figures S1–S5**), although genomospecies clusters formed at 92.5 ANI reduced the proportion of polyphyletic genomospecies within the MLST phylogeny. One of eight genomospecies (12.5%) defined at 92.5 ANI were polyphyletic based on the MLST tree, compared to 2/11 (18.2%), 3/21 (14.3%), and 4/30 (13.3%) polyphyletic genomospecies defined at 94, 95, and 96 ANI respectively (**Figure 5** and **Supplementary Figures S1–S5**).

An Adjusted, Eight-Group *panC* Framework Remains Largely Congruent With Proposed *B. cereus s.l.* Genomospecies Definitions

Another popular typing method used to assign *B. cereus s.l.* isolates to major phylogenetic groups relies on the sequence of *panC* (Guinebretiere et al., 2008, 2010). However, the seven-group *panC* framework had to be adjusted to accommodate the growing amount of *B. cereus s.l.* genomic diversity provided by WGS, as *panC* sequences assigned to Groups II, III, and VI using the seven-group typing scheme implemented in the original BTyper were polyphyletic (**Figure 4A**).

The adjusted, eight-group *panC* framework constructed here (**Figure 4B**) and implemented in BTyper3 v. 3.1.0 resolved all polyphyletic *panC* group assignments (**Figure 4**). *panC* group assignments using the adjusted, eight-group framework described here, as well as those obtained using the original seven-group framework implemented in BTyper v. 2.3.3, are available for 2,229 *B. cereus s.l.* genomes (**Table 1** and **Supplementary Tables S1, S6**). Note that group assignments using the seven-group framework implemented in the web-tool published by Guinebretiere et al. (2010) are not available, as the database is not publicly available, and the web-based method is not scalable.

However, even with an improved eight-group framework for *panC* group assignment, the *B. cereus s.l. panC* phylogeny yielded polyphyletic genomospecies, regardless of the ANI-based threshold used to define genomospecies. For seven of the eight *B. cereus s.l.* genomospecies defined at 92.5 ANI (with the exclusion of effective and proposed putative species), the *panC*

FIGURE 4 | Maximum likelihood phylogeny constructed using *panC*, extracted from 1,736 high-quality *B. cereus s.l.* genomes. Branches and tip labels are colored by **(A)** *panC* group (I–VII), assigned using the *panC* group assignment method implemented in the original BTyper v. 2.3.3 (i.e., the "legacy" *panC* group assignment method), and **(B)** adjusted *panC* group assignment (I–VIII), obtained using RhierBAPS Level 1 cluster assignments for *panC*. For all *panC* group assignments, the "foreground" *panC* group is colored (pink) and the background *panC* groups are shown in gray. Phylogenies for which the foreground *panC* group (pink) presents as polyphyletic are annotated with a pink star in the upper left corner of the panel. Phylogenies are rooted using the *panC* sequence of the *"B. manliponensis"* type strain (omitted for clarity), and branch lengths are reported in substitutions per site. IQ-TREE v. 1.6.5 was used to construct the phylogeny, using the optimal nucleotide substitution model selected using ModelFinder (i.e., the TVM + F + R4 model).

locus produced a monophyletic clade for each genomospecies (**Figures 4**, **5** and **Supplementary Figures S6, S7**). However, based on the sequence of *panC*, the *B. mosaicus* genomospecies was polyphyletic, with the *panC* sequence of *B. luti* forming a separate lineage within the *B. mosaicus panC* clade (**Figure 5** and **Supplementary Figures S6, S7**). Similarly, genomospecies defined at 94, 95, and 96 ANI produced even greater proportions of polyphyletic *panC* clusters, with 5/11 (45.5%), 8 or 9/21 (38.1 or 42.9%, depending on the phylogeny rooting method), and 8/30 (26.7%) genomospecies polyphyletic via *panC*, respectively (**Supplementary Figures S6–S15**).

rpoB Provides Lower Resolution Than *panC* for Single-Locus Sequence Typing of *B. cereus s.l.* Isolates

Another popular single-locus sequence typing method for characterizing spore-forming bacteria, including *B. cereus s.l.* isolates, relies on sequencing *rpoB*, which encodes the beta subunit of RNA polymerase (Huck et al., 2007b; Ivy et al., 2012). Among publicly available *B. cereus s.l.* isolate genomes, allelic types (ATs) assigned using the Cornell University Food Safety Laboratory and Milk Quality Improvement Program's (CUFSL/MQIP) *rpoB* allelic typing database (Carroll et al., 2017), much like STs assigned using PubMLST's seven-gene scheme (described above), were each confined to a single genomospecies at 92.5 ANI, with no AT split across genomospecies (**Supplementary Tables S1, S7**). However, fewer than 2/3 of all *B. cereus s.l.* genomes possessed a *rpoB* allele that matched a member of the database exactly (i.e., with 100% nucleotide identity and coverage; 1,425/2,231 genomes, or 63.9%). Additionally, the *B. cereus s.l. rpoB* phylogeny showcased numerous polyphyletic genomospecies, regardless of the ANI threshold at which genomospecies were defined (3/8 [37.5%], 3 or 4/11 [27.2 or 36.4%, depending on the phylogeny rooting method], 6 or 9/21 [28.6 or 42.9%, depending on the phylogeny rooting method], and 8/30 [26.7%] polyphyletic *rpoB* clades among genomospecies defined at 92.5, 94, 95, and 96 ANI, respectively; **Figure 5** and **Supplementary Figures S16–S23**). Thus, at the present time, *rpoB* alleles that match a member of the

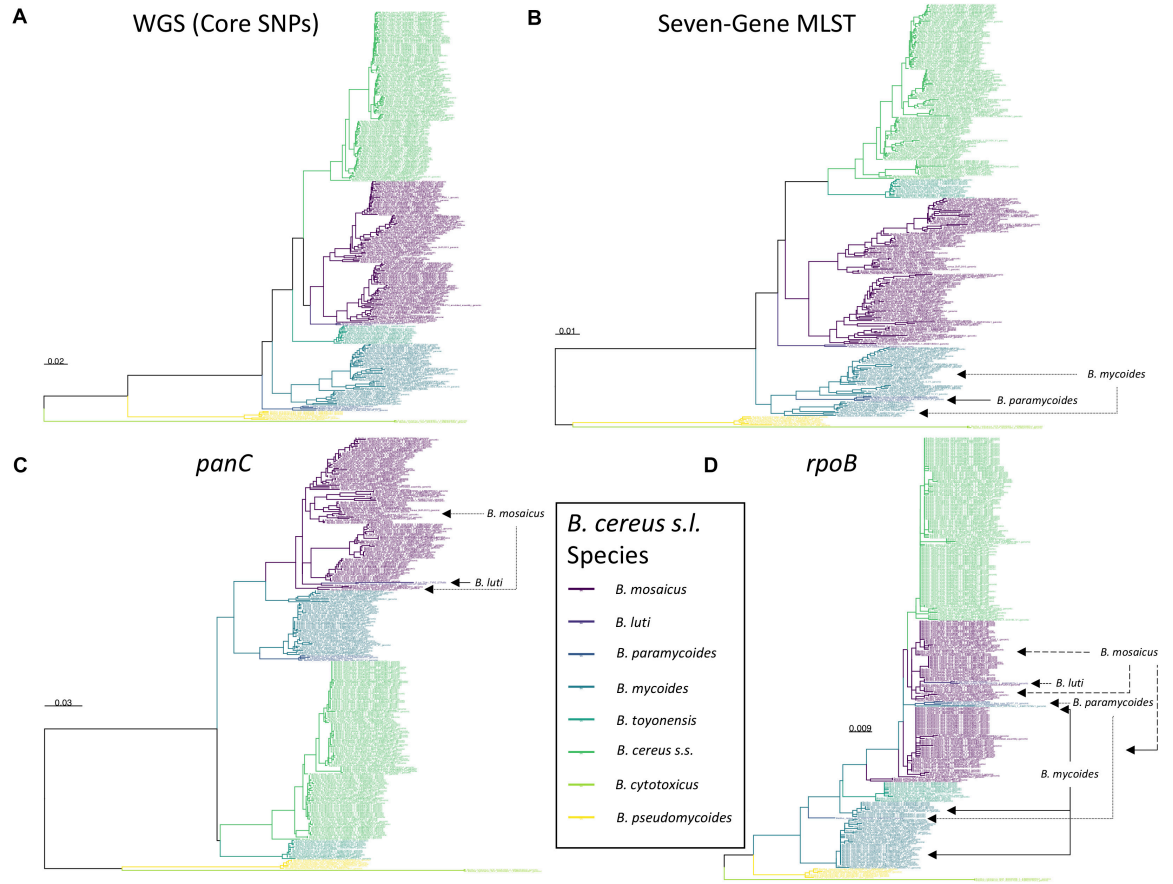

FIGURE 5 | Maximum likelihood phylogenies constructed using **(A)** genome-wide core SNPs (WGS), **(B)** seven concatenated multi-locus sequence typing (MLST) genes, **(C)** *panC,* and **(D)** *rpoB* identified among 313 high-quality *B. cereus s.l.* medoid genomes identified at a 99 average nucleotide identity (ANI) threshold. Branches and tip labels are colored by ANI-based genomospecies assignment using the proposed *B. cereus s.l.* taxonomic framework (i.e., eight genomospecies assigned using medoid genomes and a 92.5 ANI threshold). Polyphyletic genomospecies and the genomospecies interspersed among them are annotated with arrows. Phylogenies are rooted at the midpoint, and branch lengths are reported in substitutions per site. Genomospecies were assigned using BTyper3 v. 3.1.0 and FastANI v. 1.0. For the WGS tree **(A)**, core SNPs were identified among all 313 *B. cereus s.l.* genomes using kSNP3 v. 3.92 and the optimal *k*-mer size reported by Kchooser (k = 19). For the MLST, *panC*, and *rpoB* trees (**B–D**, respectively), BTyper v. 2.3.3 was used to extract all loci from the set of 313 *B. cereus s.l.* genomes, and MAFFT v. 7.453-with-extensions was used to construct an alignment for each locus. For each alignment, IQ-TREE v. 1.5.4 **(A)** and 1.6.5 **(B–D)** was used to construct a phylogeny, using either the GTR + G + ASC nucleotide substitution model **(A)**, or the optimal model selected using ModelFinder **(B–D)**.

CUFSL/MQIP *rpoB* allelic typing database exactly (i.e., with 100% identity and coverage) can be used for *B. cereus s.l.* genomospecies assignment. However, due to the low resolution and high frequency of polyphyletic genomospecies in the *rpoB* phylogeny, *rpoB* alleles that do not match a member of the database exactly should not be used for genomospecies assignment and should be interpreted with caution.

Numerous Single Loci Mirror the Topology of *B. cereus s.l.* and May Provide Improved Resolution for Single- and/or Multi-Locus Sequence Typing

A total of 1,719 single-copy loci were present among 313 high-quality *B. cereus s.l.* medoid genomes identified at 99 ANI (this was done to remove highly similar genomes and reduce the search space). After alignment, 255 of the 1,719 loci (i) produced an alignment that did not include any gap characters among at least 90% of its sites and (ii) contained a continuous sequence, uninterrupted by gaps, which covered at least 90% of total sites within the alignment, (iii) were present in a single copy in all 1,741 high-quality *B. cereus s.l.* genomes, sharing $\geq$ 90% nucleotide identity and coverage with at least one of the 313 alleles extracted from each of the 313 99 ANI medoid genomes, and (iv) produced a maximum likelihood phylogeny which mirrored the WGS phylogeny (Kendall–Colijn $P < 0.05$ after a Bonferroni correction; **Supplementary Table S2**). The resulting 255 single-copy core loci spanned a wide array of functions and were predicted to be involved in a diverse range of biological processes, including sporulation and response to stress (**Figure 6**, **Supplementary Figure S24**, and **Supplementary Table S2**). Future single- and/or MLST schemes querying one or more of these loci may improve taxonomic assignment; however, additional work is needed to validate and

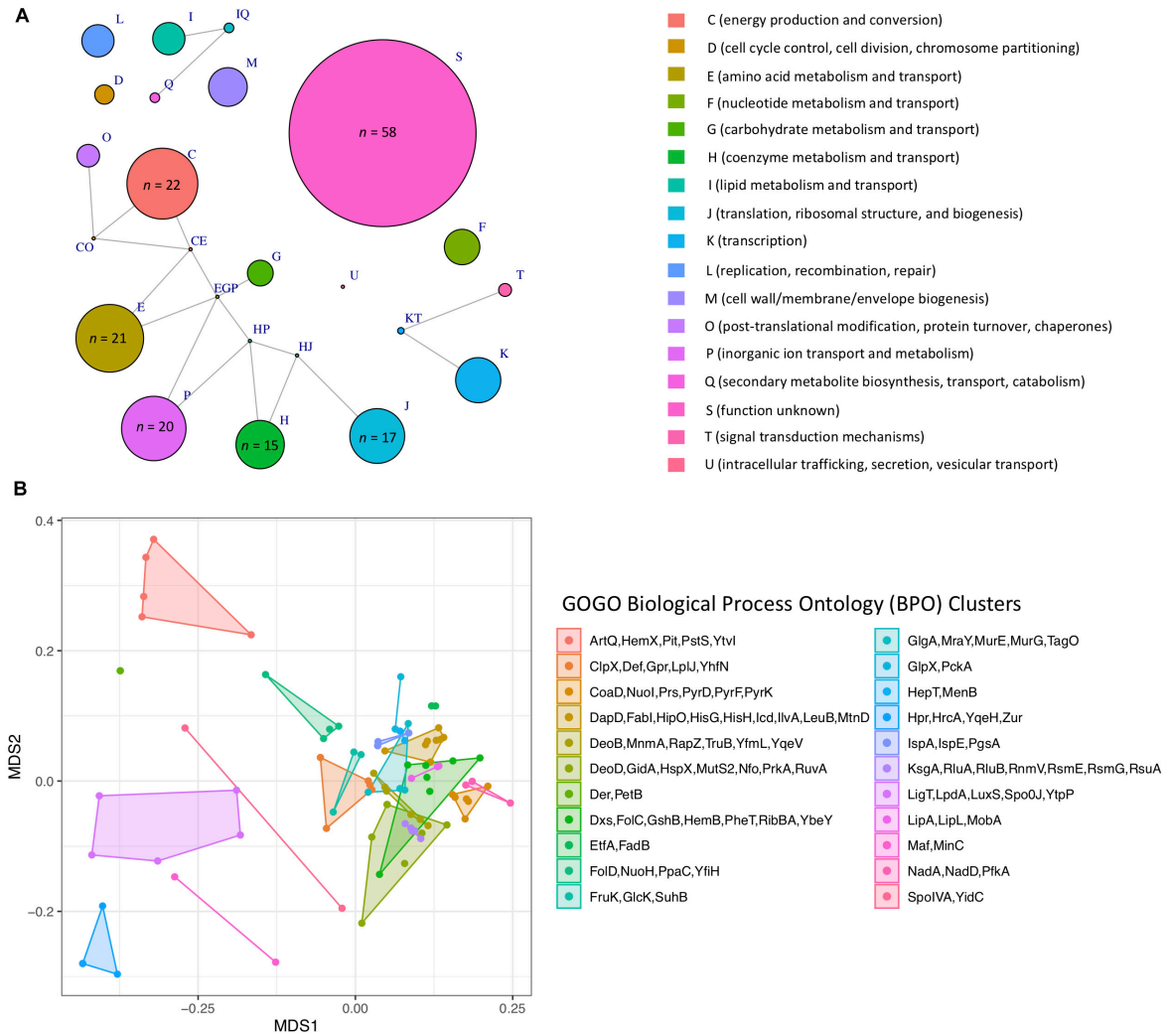

FIGURE 6 | (A) Network of Clusters of Orthologous Groups (COG) functional categories assigned to 255 single-copy core genes that topologically mirror the *B. cereus s.l.* whole-genome phylogeny (Kendall–Colijn $P < 0.05$ after a Bonferroni correction). Each node corresponds to a COG functional category/group of functional categories assigned to one or more genes. Node size corresponds to the number of genes (out of 255 possible genes) assigned to a functional category/group of functional categories, ranging from one to 58 (for S, function unknown). Edges connect nodes that share one or more functional categories. Nodes of functional categories assigned to 15 or more genes are annotated with a text label denoting the number of genes assigned to the respective functional category. **(B)** Results of non-metric multidimensional scaling (NMDS) performed using pairwise semantic/functional dissimilarities calculated between 94 single-copy core genes based on their assigned Gene Ontology (GO) Biological Process Ontology (BPO) terms. Points represent individual genes, while shaded regions and convex hulls correspond to clusters of genes identified by GOGO, based on their BPO similarities. For a complete list of annotations associated with each of the 255 single-copy core genes, see **Supplementary Table S2**. For NMDS plots constructed using Cellular Component Ontology (CCO) and Molecular Function Ontology (MFO) dissimilarities, see **Supplementary Figure S24**.

assess the robustness of these loci as taxonomic markers in an experimental setting.

The Adjusted, Eight-Group *panC* Framework Captures Genomic Heterogeneity of Anthrax-Causing "*B. cereus*"

The set of 1,741 high-quality *B. cereus s.l.* genomes was queried for *B. cereus s.l.* virulence factors with known associations to anthrax (Okinaka et al., 1999; Candela and Fouet, 2006; Oh et al., 2011), emetic (Ehling-Schulz et al., 2006, 2015), and diarrheal illnesses (Schoeni and Wong, 2005; Stenfors Arnesen et al., 2008; Fagerlund et al., 2010; Senesi and Ghelardi, 2010) using amino acid identity and coverage thresholds of 70% and 80%, respectively (**Figure 3**). Using the proposed genomospecies-subspecies-biovar taxonomy and operon/cluster-level groupings for virulence factors (where applicable), cereulide synthetase-encoding *cesABCD* were detected in (i) the *B. mosaicus* and *B. mycoides* genomospecies and (ii) *panC* Group III and VI, respectively, as described previously (Guinebretiere et al., 2008, 2010; Carroll et al., 2017, 2020b;

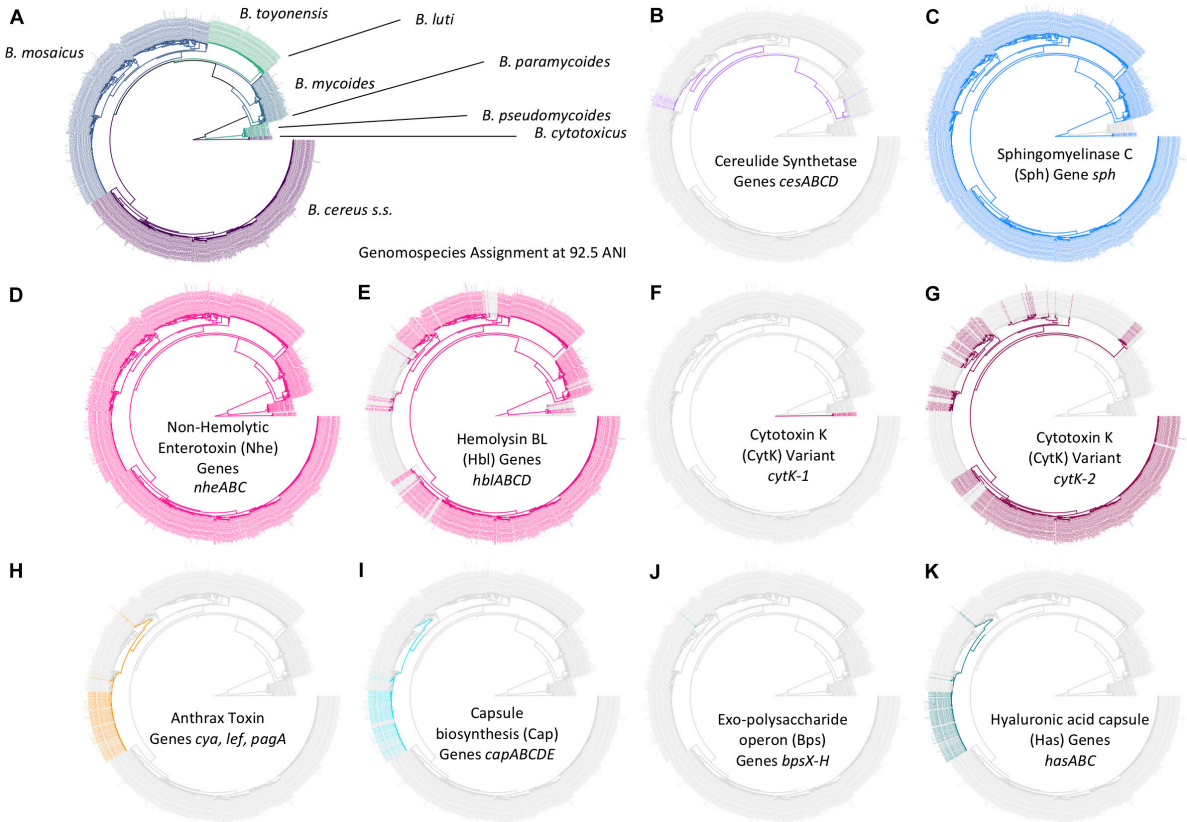

FIGURE 7 | Distribution of selected *B. cereus s.l.* virulence factors within the *B. cereus s.l.* phylogeny (n = 1,741). Tip labels and branches within the phylogeny are colored by **(A)** *B. cereus s.l.* genomospecies, assigned using medoid genomes obtained at a 92.5 ANI threshold, and **(B–K)** presence and absence of the denoted *B. cereus s.l.* virulence factor (colored and gray tip labels, respectively). Virulence factors were detected using BTyper v. 3.1.0, with minimum amino acid identity and coverage thresholds of 70 and 80%, respectively, and a maximum E-value threshold of 1E-5. A virulence factor was considered to be present in a genome if all genes comprising the virulence factor were detected at the aforementioned thresholds; likewise, if one or more genes comprising a virulence factor were not detected at the given thresholds, the virulence factor was considered to be absent. The phylogeny was constructed using core SNPs identified in 79 single-copy orthologous gene clusters present among all 2,231 *B. cereus s.l.* genomes available in NCBI's RefSeq database (accessed 19 November 2018; see Carroll et al., 2020b, for detailed methods) (Carroll et al., 2020b). The type strain of "*B. manliponensis*" (i.e., the most distantly related member of the group) was treated as an outgroup on which the phylogeny was rooted. Tips representing genomes that (i) did not meet the quality thresholds and/or (ii) were not assigned to one of eight published genomospecies (i.e., genomes of unpublished, proposed, or effective *B. cereus s.l.* species) were omitted.

Carroll and Wiedmann, 2020) and regardless of whether the legacy seven-group or adjusted eight-group *panC* typing schemes were used (**Figure 7** and **Supplementary Table S1**).

Anthrax toxin genes and anthrax-associated capsule-encoding operons *cap*, *has*, and *bps* were detected in their entirety in the *B. mosaicus* genomospecies alone (**Figure 7** and **Supplementary Table S1**). Using the legacy, seven-group *panC* group assignment scheme implemented in the original BTyper (i.e., BTyper v. 2.3.3), all anthrax-associated virulence factors were confined to *panC* Group III; however, using the adjusted, eight-group framework, some anthrax-causing strains were assigned to Group II (**Figure 7** and **Supplementary Table S1**). All anthrax-causing strains that belonged to the non-motile, non-hemolytic (Tallent et al., 2012, 2019) highly similar (≥99.9 ANI) (Jain et al., 2018; Carroll et al., 2020b) lineage commonly associated with anthrax disease (known as species *B. anthracis*; using the proposed taxonomy, *B. anthracis* biovar Anthracis or *B. mosaicus* subsp. *anthracis* biovar Anthracis using subspecies and full notation, respectively) remained in *panC* Group III (**Supplementary Table S1**). However, the eight-group *panC* framework was able to capture genomic differences between anthrax-causing strains with phenotypic characteristics resembling "*B. cereus*" (e.g., motility, hemolysis; see **Supplementary Table S1** here or **Supplementary Table S5** of Carroll et al., 2020b, for a list of strains). Known previously as anthrax-causing "*B. cereus*" or "*B. cereus*" biovar Anthracis, among other names (using the proposed 2020 taxonomy, *B. mosaicus* biovar Anthracis), these strains could be partitioned into two major lineages: one that more closely resembled *B. anthracis* and one that more closely resembled *B. tropicus* using ANI-based comparisons to species type strains that existed in 2019 (Carroll et al., 2020b). These anthrax-causing "*B. cereus*" lineages were assigned to *panC* Groups III and II using the adjusted, eight-group *panC* framework developed here, respectively (**Supplementary Table S1**).

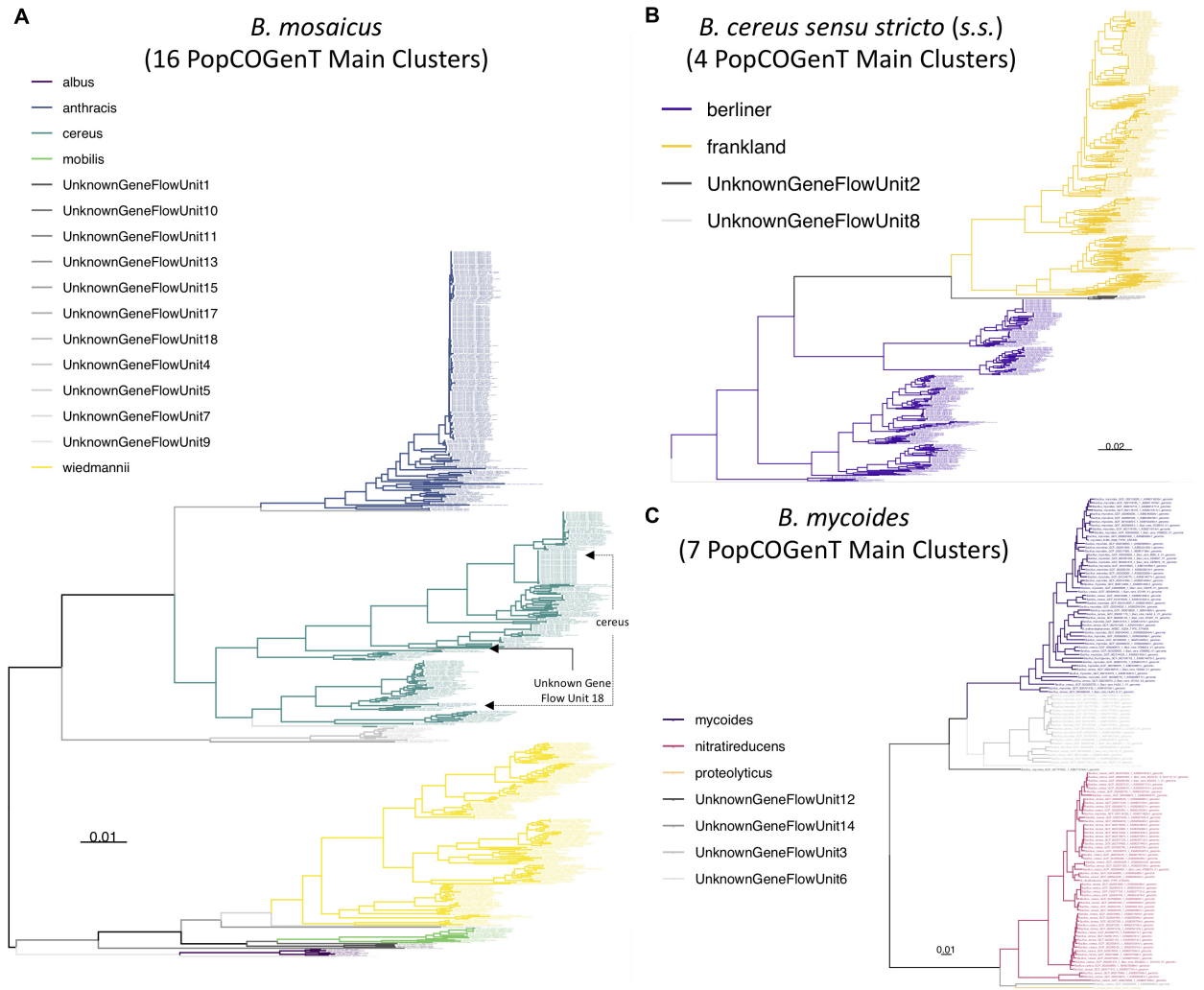

FIGURE 8 | Maximum likelihood phylogenies constructed using genome-wide core SNPs identified among all high-quality genomes assigned to each of the **(A)** *B. mosaicus*, **(B)** *B. cereus sensu stricto* (*s.s.*), and **(C)** *B. mycoides* genomospecies delineated at a 92.5 ANI threshold. Branches and tip labels are colored by pseudo-gene flow unit assignment using the pseudo-gene flow unit assignment algorithm implemented in BTyper3 v. 3.1.0 (**Figures 1**, **2**). Only genomes that fell within the observed ANI boundary for each pseudo-gene flow unit are shown. Arrows are used to annotate polyphyletic pseudo-gene flow units derived from "true" gene flow units that also presented as polyphyletic (**Supplementary Figure S25**); the pseudo-gene flow units interspersed among them are additionally annotated with arrows. Phylogenies are rooted at the midpoint, and branch lengths are reported in substitutions per site. Genomospecies and pseudo-gene flow units were assigned using BTyper3 v. 3.1.0 and FastANI v. 1.0. For each phylogeny, core SNPs were identified among all high-quality genomes assigned to the genomospecies using kSNP3 v. 3.92 and the optimal *k*-mer size reported by Kchooser (*k* = 19 or 21). For each core SNP alignment, IQ-TREE v. 1.5.4 was used to construct a phylogeny, using the GTR + G + ASC nucleotide substitution model.

Diarrheal enterotoxin-encoding genes were widespread throughout the *B. cereus s.l.* phylogeny (**Figure 7** and **Supplementary Table S1**), as many others have noted before (Guinebretiere et al., 2008, 2010; Stenfors Arnesen et al., 2008; Kovac et al., 2016; Carroll et al., 2017). Nhe-encoding *nheABC* were detected in nearly all genomes (1,731/1,741 genomes, 99.4%; **Figure 7** and **Supplementary Table S1**). Hbl-encoding *hblABCD* were detected in one or more members of all genomospecies except *B. cytotoxicus* and *B. luti* (**Figure 7** and **Supplementary Table S1**). Variant 2 of CytK-encoding *cytK* (i.e., *cytK-2*) was identified in *B. cereus s.s.* (Group IV), *B. mosaicus* (Groups II and III), and *B. toyonensis* (Group V); variant 1 (*cytK-1*) was exclusive to *B. cytotoxicus*, as noted previously (Fagerlund et al., 2004; Guinebretiere et al., 2006; Carroll et al., 2017; Stevens et al., 2019).

A Method Querying Recent Gene Flow Identifies Multiple Major Gene Flow Units Among the *B. cereus s.s.*, *B. mosaicus*, *B. mycoides*, and *B. toyonensis* Genomospecies

A recently proposed method for delineating microbial gene flow units using recent gene flow (referred to hereafter as the

"populations as clusters of gene transfer," or "PopCOGenT," method) (Arevalo et al., 2019) was applied to the set of 313 high-quality *B. cereus s.l.* medoid genomes identified at 99 ANI. The PopCOGenT method identified a total of 33 "main clusters," or gene flow units that attempt to mimic the classical species definition used for animals and plants (**Table 2**). Minimum ANI values shared between isolates assigned to the same gene flow unit ranged from 94.7 to 98.9 ANI for clusters containing more than one isolate (**Table 2**). A "pseudo-gene flow unit" assignment method was implemented in BTyper3 v. 3.1.0, in which ANI values are calculated between a query genome and the medoid genomes of the 33 PopCOGenT gene flow units using FastANI; if the query genome shares an ANI value with one of the gene flow unit medoid genomes that is greater than or equal to the previously observed ANI boundary for the gene flow unit, it is assigned to that particular pseudo-gene flow unit (**Figures 1**, **2** and **Table 2**). This pseudo-gene flow unit assignment method was applied to all 2,231 *B. cereus s.l.* genomes (**Table 2** and **Supplementary Table S1**), and was found to yield pseudo-gene flow units that were each encompassed within a single genomospecies and *panC* group (using the adjusted eight-group *panC* scheme developed here), with no pseudo-gene flow units split across multiple genomospecies/*panC* groups (**Table 2**). PopCOGenT identified multiple gene flow units among the *B. cereus s.s.*, *B. mosaicus, B. mycoides,* and *B. toyonensis* genomospecies delineated at 92.5 ANI (n = 4, 16, 7, and 2 main clusters, respectively; **Figure 8** and **Supplementary Figures S25–S32**).

DISCUSSION

The Proposed *B. cereus s.l.* Taxonomy Is Backward-Compatible With *B. cereus s.l.* Species Defined Using Historical ANI-Based Species Thresholds

ANI-based methods have been used to define 12 *B. cereus s.l.* species prior to 2020: *B. cytotoxicus* and *B. toyonensis*, each proposed as novel species in 2013 (Guinebretiere et al., 2013; Jimenez et al., 2013), *B. wiedmannii* (proposed as a novel species in 2016) (Miller et al., 2016), and nine species (*B. albus, B. luti, B. mobilis, B. nitratireducens, B. pacificus, B. paranthracis*, *B. paramycoides, B. proteolyticus,* and *B. tropicus*) proposed in 2017 (Liu et al., 2017). However, the lack of a standardized ANI-based genomospecies threshold for defining *B. cereus s.l.* genomospecies has led to confusion regarding how *B. cereus s.l.* species should be delineated. *B. toyonensis* and the nine species proposed in 2017, for example, were defined using genomospecies thresholds of 94 and 96 ANI, respectively (Jimenez et al., 2013; Liu et al., 2017). The descriptions of *B. cytotoxicus* and *B. wiedmannii* as novel species each explicitly state that a 95 ANI threshold was used (Guinebretiere et al., 2013; Miller et al., 2016); however, the *B. wiedmannii* type strain genome shared a much higher degree of similarity with the type strain genome of its neighboring species than did *B. cytotoxicus* (Miller et al., 2016). As such, choice of ANI-based genomospecies threshold can affect which *B. cereus s.l.* strains belong to which genomospecies, and may even produce overlapping genomospecies in which a genome can belong to more than one genomospecies (Carroll et al., 2020b).

The proposed *B. cereus s.l.* taxonomy (Carroll et al., 2020b) provides a standardized genomospecies threshold of 92.5 ANI, which has been shown to yield non-overlapping, monophyletic *B. cereus s.l.* genomospecies. However, the practice of assigning *B. cereus s.l.* isolates to genomospecies using species type strain genomes and historical species thresholds (i.e., 94–96 ANI) has been important for whole-genome characterization for *B. cereus s.l.* strains, including those responsible for illnesses and/or outbreaks (Lazarte et al., 2018; Bukharin et al., 2019; Carroll et al., 2019). Here, we show that all 18 published *B. cereus s.l.* genomospecies defined prior to 2020 are safely integrated into the proposed *B. cereus s.l.* taxonomy without polyphyly, regardless of whether a 94, 95, or 96 ANI genomospecies threshold was used to delineate species relative to type strain genomes.

Single- and Multi-Locus Sequence Typing Methods Can Be Used to Assign *B. cereus s.l.* Isolates to Species Within the Proposed Taxonomy

Single- and multi-locus sequence typing approaches have been—and continue to be—important methods for classifying *B. cereus s.l.* isolates into phylogenetic units. They have been used to characterize *B. cereus s.l.* strains associated with illnesses and outbreaks (Cardazzo et al., 2008; Glasset et al., 2016; Akamatsu et al., 2019; Carroll et al., 2019), strains isolated from food and food processing environments (Huck et al., 2007a; Thorsen et al., 2015; Kindle et al., 2019; Ozdemir and Arslan, 2019; Zhuang et al., 2019; Zhao et al., 2020), and strains with industrial applications (e.g., biopesticide strains) (Johler et al., 2018). Additionally, STs and ATs assigned using these approaches have been used to construct frameworks for predicting the risk that a particular *B. cereus s.l.* strain poses to food safety, public health, and food spoilage (Guinebretiere et al., 2010; Rigaux et al., 2013; Buehler et al., 2018; Miller et al., 2018; Webb et al., 2019). It is thus important to ensure that the proposed standardized taxonomy for *B. cereus s.l.* remains congruent with widely used sequence typing approaches.

Here, we assessed the congruency of three popular single- and multi-locus sequence typing schemes for *B. cereus s.l.* with proposed genomospecies definitions: (i) the PubMLST seven-gene MLST scheme for *B. cereus s.l.* (Jolley and Maiden, 2010; Jolley et al., 2018), (ii) the seven-group *panC* typing scheme developed by Guinebretiere et al. (2008, 2010) as implemented in the original BTyper (Carroll et al., 2017), and (iii) the CUFSL/MQIP *rpoB* allelic typing scheme used for characterizing spore-forming bacteria, including members of *B. cereus s.l.* (Durak et al., 2006; Huck et al., 2007a; Ivy et al., 2012; Buehler et al., 2018). STs and ATs assigned using MLST and *rpoB* allelic typing, respectively, were each contained within a single genomospecies at 92.5 ANI. Consequently, past studies employing these methods can be easily interpreted within the proposed taxonomic framework for the group. Additionally,

TABLE 2 | Gene flow units delineated using recent gene flow[a].

Cluster #	Encompassing species	Minimum ANI value[b]	Notable members within minimum ANI bound (relative to PopCOGenT medoid)	*panC* group[c]	Proposed gene flow unit name
0	*B. mosaicus*	97.9	*B. albus*^{T}	II	*albus*
1	*B. luti*	96.6	*B. luti*^{T}	II	*luti*
2	*B. mosaicus*	96.8	*B. mobilis*^{T}	II	*mobilis*
3	*B. paramycoides*	97.1	*B. paramycoides*^{T}	VI	*paramycoides*
4	*B. toyonensis*	97.8	*B. toyonensis*^{T}	V	*toyonensis*
5	*B. mosaicus*	96.7	*B. anthracis* str. Ames	III	*anthracis*
6	*B. mosaicus*	94.7	Emetic reference *B. cereus* str. AH187, *B. paranthracis*^{T}, *B. pacificus*^{T}, *B. tropicus*^{T}	III	*cereus*
7	*B. cereus s.s.*	96.0	*B. cereus s.s.* ATCC 14579^T	IV	*frankland*
8	*B. mosaicus*	98.0		II	Unknown Unit 1
9	*B. mycoides*	96.1	*B. mycoides*^{T}, *B. weihenstephanensis*^{T}	VI	*mycoides*
10	*B. cereus s.s.*	98.7		IV	Unknown Unit 2
11	*B. mycoides*	96.9		VI	Unknown Unit 3
12	*B. cereus s.s.*	95.6	*B. thuringiensis* serovar berliner ATCC 10792^T	IV	*berliner*
13	*B. mycoides*	95.3	*B. nitratireducens*^{T}	VI	*nitratireducens*
14	*B. mosaicus*	95.7	*B. wiedmannii*^{T}	II	*wiedmannii*
15	*B. mosaicus*	97.3		II	Unknown Unit 4
16	*B. cytotoxicus*	98.9	*B. cytotoxicus*^{T}	VII	*cytotoxicus*
17	*B. pseudomycoides*	95.9	*B. pseudomycoides*^{T}	I	*pseudomycoides*
18	*B. mosaicus*	100.0		II	Unknown Unit 5
19	*B. mycoides*	100.0		VI	Unknown Unit 6
20	*B. mosaicus*	100.0		II	Unknown Unit 7
21	*B. cereus s.s.*	100.0		IV	Unknown Unit 8
22	*B. mosaicus*	100.0		II	Unknown Unit 9
23	*B. mosaicus*	100.0		II	Unknown Unit 10
24	*B. mosaicus*	100.0		II	Unknown Unit 11
25	*B. mycoides*	100.0		VI	Unknown Unit 12
26	*B. mosaicus*	100.0		II	Unknown Unit 13
27	*B. mycoides*	100.0	*B. proteolyticus*^{T}	VIII	*proteolyticus*
28	*B. mycoides*	100.0		VIII	Unknown Unit 14
29	*B. mosaicus*	100.0		II	Unknown Unit 15
30	*B. toyonensis*	100.0		V	Unknown Unit 16
31	*B. mosaicus*	100.0		II	Unknown Unit 17
32	*B. mosaicus*	100.0		II	Unknown Unit 18

[a] *See* ***Supplementary Tables S5, S7*** *for sequence types and rpoB allelic types associated with each taxonomic group.* [b] *Minimum average nucleotide identity (ANI) value for the cluster;* [c] *panC group assignment using the adjusted eight-group framework described here.*

six of eight *panC* groups assigned using the adjusted eight-group framework developed here were each contained within a single genomospecies delineated at 92.5 ANI (i.e., all but Group II, which contained *B. mosaicus* and *B. luti*, and Group VI, which contained *B. mycoides* and *B. paramycoides*). Thus, unlike genomospecies delineated at higher thresholds (i.e., 94–96 ANI), *panC* can be used for assignment of most *B. cereus s.l.* genomospecies delineated at 92.5 ANI.

MLST, *panC* group assignment, and *rpoB* allelic typing will likely remain extremely valuable for characterizing *B. cereus s.l.* isolates, as all three approaches remain largely congruent with *B. cereus s.l.* genomospecies defined at 92.5 ANI. However, all three typing methods produced at least one polyphyletic genomospecies among genomospecies defined at 92.5 ANI. Higher, historical genomospecies thresholds (i.e., 94, 95, and 96 ANI) showcased even higher proportions of polyphyly within the MLST and *panC* phylogenies. This observation is particularly important for *panC* group assignment, as *panC* may not be able to differentiate between some members of *B. mosaicus* and *B. luti* (each assigned to *panC* Group II) with adequate resolution. In addition to assessing the congruency of proposed typing methods, we used a computational approach to identify putative loci that may better capture the topology of the whole-genome *B. cereus s.l.* phylogeny. While typing schemes that incorporate these loci still need to be validated in an experimental setting, future single-locus sequence typing methods using loci that mirror the "true" topology of *B. cereus s.l.* may improve sequence typing efforts.

A Rapid, Scalable ANI-Based Method Can Be Used to Assign Genomes to Pseudo-Gene Flow Units Identified Among *B. cereus s.l.* Genomospecies

ANI-based methods have become the gold standard for bacterial taxonomy in the WGS era (Richter and Rossello-Mora, 2009), as they conceptually mirror DNA-DNA hybridization and implicitly account for the fluidity that accompanies bacterial genomes

(Jain et al., 2018). However, the concept of the bacterial "species" has been, and remains, controversial, as the promiscuous genetic exchange that occurs among prokaryotes can obscure population boundaries (Hanage et al., 2005; Rocha, 2018; Arevalo et al., 2019). Recently, Arevalo et al. (2019) outlined a method that attempts to delineate microbial gene flow units and the populations within them using a metric based on recent gene flow. The resulting gene flow units identified among bacterial genomes are proposed to mimic the classical species definition used for plants and animals (i.e., interbreeding units separated by reproductive barriers) (Huxley, 1943; Arevalo et al., 2019). Here, we used PopCOGenT to characterize a subset of isolates that capture genomic diversity across *B. cereus s.l.*, and we identified 33 main gene flow units among *B. cereus s.l.* isolates assigned to known genomospecies.

While the PopCOGenT method attempts to apply classical definitions of species developed with higher organisms in mind to microbes, we propose to maintain ANI-based *B. cereus s.l.* genomospecies definitions (i.e., ANI-based genomospecies clusters formed using medoid genomes obtained at a 92.5 ANI breakpoint) due to (i) the speed, scalability, portability, and accessibility of the ANI algorithm, and (ii) the accessibility and backward-compatibility of the eight-genomospecies *B. cereus s.l.* taxonomic framework, as demonstrated in this study. ANI is fast and can readily scale to large numbers (e.g., tens of thousands) of bacterial genomes (Jain et al., 2018), traits that will become increasingly important as more *B. cereus s.l.* genomes are sequenced. In addition to speed and scalability, ANI is a well-understood algorithm implemented in many easily accessible tools, including command-line tools (e.g., FastANI, pyani, OrthoANI), desktop applications (e.g., JSpecies, OrthoANI), and web-based tools (e.g., JSpeciesWS, MiGA, OrthoANIu) (Goris et al., 2007; Richter and Rossello-Mora, 2009; Lee et al., 2016; Pritchard et al., 2016; Richter et al., 2016; Yoon et al., 2017; Jain et al., 2018; Rodriguez et al., 2018). Finally, the gene flow units identified using the PopCOgenT method in the present study were not congruent with historical ANI-based genomospecies assignment methods used for *B. cereus s.l.* Genomospecies defined at historical ANI thresholds are not readily integrated into the gene flow units identified via the PopCOgenT method, as the ANI boundaries for PopCOgenT gene flow units vary (**Table 2**).

Despite its infancy and current limitations, the PopCOgenT framework provides an interesting departure from a one-threshold-fits-all ANI-based taxonomy. Here, we implemented a "pseudo-gene flow unit" method in BTyper3 v. 3.1.0 that can be used to assign a user's genome of interest to a pseudo-gene flow unit using the set of 33 PopCOgenT gene flow unit medoid genomes, the pairwise ANI values calculated within PopCOgenT gene flow units, and FastANI. However, it is essential to note the limitations of the pseudo-gene flow unit assignment method implemented in BTyper3. First and foremost, ANI and the methods employed by PopCOgenT are fundamentally and conceptually different; the pseudo-gene flow unit assignment method described here does not infer recent gene flow, nor does it use PopCOgenT or any of its metrics. Thus, the pseudo-gene flow unit assignment method cannot be used to construct true gene flow units for *B. cereus s.l.* Secondly, to increase the speed of PopCOGenT, we reduced *B. cereus s.l.* to a set of 313 representative genomes that encompassed the diversity of the species complex; genomes that shared ≥99 ANI with one or more genomes in this representative set were omitted (i.e., 1,428 of 1,741 high-quality genomes were omitted; 82.0%). Consequently, gene flow among most closely related lineages that shared ≥99 ANI with each other was not assessed, as it was thereby assumed that highly similar genomes that shared ≥99 ANI with each other belonged to the same PopCOGenT "main cluster" (i.e., "species"). It is possible that the inclusion of these highly similar genomes would have resulted in the discovery of additional gene flow units, or perhaps changes in existing ones, and future studies are needed to assess and refine this. However, the pseudo-gene flow unit assignment approach described here allows researchers to rapidly identify the most similar medoid genome of true gene flow units identified within *B. cereus s.l.* Results should not be interpreted as an assessment of recent gene flow, but rather as a higher-resolution phylogenomic clade assignment, similar to how one might use MLST for delineation of lineages within species. We anticipate that our rapid method will be valuable to researchers who desire greater resolution than what is provided at the genomospecies level, particularly when querying diverse *B. cereus s.l.* genomospecies that comprise multiple major clades (e.g., *B. mosaicus*, *B. mycoides*, *B. cereus s.s.*).

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

LC performed all computational analyses. LC, RC, and JK designed the study and co-wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by USDA NIFA Hatch Appropriations under project no. PEN04646 and accession no. 1015787, and the USDA NIFA grant GRANT12686965.

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at BioRxiv (Carroll et al., 2020a).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.580691/full#supplementary-material

REFERENCES

Acevedo, M. M., Carroll, L. M., Mukherjee, M., Mills, E., Xiaoli, L., Dudley, E. G., et al. (2019). *Bacillus clarus* sp. nov. is a new *Bacillus cereus* group species isolated from soil. *bioRxiv* [Preprint]. doi: 10.1101/508077

Akamatsu, R., Suzuki, M., Okinaka, K., Sasahara, T., Yamane, K., Suzuki, S., et al. (2019). Novel Sequence Type in *Bacillus cereus* strains associated with nosocomial infections and bacteremia, Japan. *Emerg. Infect. Dis* 25, 883–890. doi: 10.3201/eid2505.171890

Angiuoli, S. V., and Salzberg, S. L. (2011). Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* 27, 334–342. doi: 10.1093/bioinformatics/btq665

Antonation, K. S., Grutzmacher, K., Dupke, S., Mabon, P., Zimmermann, F., Lankester, F., et al. (2016). *Bacillus cereus* Biovar anthracis causing anthrax in sub-saharan Africa-chromosomal monophyly and broad geographic distribution. *PLoS Negl. Trop. Dis.* 10:e0004923. doi: 10.1371/journal.pntd.0004923

Arevalo, P., Vaninsberghe, D., Elsherbini, J., Gore, J., and Polz, M. F. (2019). A reverse ecology approach based on a biological definition of microbial populations. *Cell* 178:e814. doi: 10.1016/j.cell.2019.06.033

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Gene Ontol. Consortium. Nat Genet.* 25, 25–29. doi: 10.1038/75556

Avashia, S. B., Riggins, W. S., Lindley, C., Hoffmaster, A., Drumgoole, R., Nekomoto, T., et al. (2007). Fatal pneumonia among metalworkers due to inhalation exposure to *Bacillus cereus* Containing Bacillus anthracis toxin genes. *Clin. Infect. Dis.* 44, 414–416. doi: 10.1086/510429

Brezillon, C., Haustant, M., Dupke, S., Corre, J. P., Lander, A., Franz, T., et al. (2015). Capsules, toxins and AtxA as virulence factors of emerging *Bacillus cereus* biovar anthracis. *PLoS Negl. Trop. Dis.* 9:e0003455. doi: 10.1371/journal.pntd.0003455

Buehler, A. J., Martin, N. H., Boor, K. J., and Wiedmann, M. (2018). Psychrotolerant spore-former growth characterization for the development of a dairy spoilage predictive model. *J. Dairy Sci.* 101, 6964–6981. doi: 10.3168/jds.2018-14501

Bukharin, O. V., Perunova, N. B., Andryuschenko, S. V., Ivanova, E. V., Bondarenko, T. A., and Chainikova, I. N. (2019). Genome Sequence Announcement of Bacillus paranthracis Strain ICIS-279, Isolated from Human Intestine. *Microbiol. Resour. Announc.* 8:e00662-19. doi: 10.1128/MRA.00662-19

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinform.* 10:421. doi: 10.1186/1471-2105-10-421

Candela, T., and Fouet, A. (2006). Poly-gamma-glutamate in bacteria. *Mol. Microbiol.* 60, 1091–1098. doi: 10.1111/j.1365-2958.2006.05179.x

Cardazzo, B., Negrisolo, E., Carraro, L., Alberghini, L., Patarnello, T., and Giaccone, V. (2008). Multiple-locus sequence typing and analysis of toxin genes in *Bacillus cereus* food-borne isolates. *Appl. Environ. Microbiol.* 74, 850–860. doi: 10.1128/AEM.01495-07

Carroll, L. M., Cheng, R. A., and Kovac, J. (2020a). No assembly required: using BTyper3 to assess the congruency of a proposed taxonomic framework for the *Bacillus cereus* group with historical typing methods. *bioRxiv* [Preprint]. doi: 10.1101/2020.06.28.175992

Carroll, L. M., Kovac, J., Miller, R. A., and 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 data. *Appl. Environ. Microbiol.* 83:e01096-17. doi: 10.1128/AEM.01096-17

Carroll, L. M., and Wiedmann, M. (2020). Cereulide synthetase acquisition and loss events within the evolutionary history of Group III *Bacillus cereus* sensu lato facilitate the transition between emetic and diarrheal foodborne pathogen. *bioRxiv* [Preprint]. doi: 10.1128/mBio.01263-20

Carroll, L. M., Wiedmann, M., and Kovac, J. (2020b). Proposal of a taxonomic nomenclature for the *Bacillus cereus* group which reconciles genomic definitions of bacterial species with clinical and industrial phenotypes. *mBio* 11:e00034-20. doi: 10.1128/mBio.00034-20

Carroll, L. M., Wiedmann, M., Mukherjee, M., Nicholas, D. C., Mingle, L. A., Dumas, N. B., et al. (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.* 10:144. doi: 10.3389/fmicb.2019.00144

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *Int. J., Complex Syst.* 1695, 1-9.

Durak, M. Z., Fromm, H. I., Huck, J. R., Zadoks, R. N., and Boor, K. J. (2006). Development of molecular typing methods for Bacillus spp. *and Paenibacillus* spp. Isolated from fluid milk products. *J. Food Sci.* 71, M50–M56. doi: 10.1111/j.1365-2621.2006.tb08907.x

Ehling-Schulz, M., Frenzel, E., and Gohar, M. (2015). Food-bacteria interplay: pathometabolism of emetic *Bacillus cereus*. *Front. Microbiol.* 6:704. doi: 10.3389/fmicb.2015.00704

Ehling-Schulz, M., Fricker, M., Grallert, H., Rieck, P., Wagner, M., and 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.* 6:20. doi: 10.1186/1471-2180-6-20

Ehling-Schulz, M., Lereclus, D., and Koehler, T. M. (2019). The *Bacillus cereus* Group: Bacillus Species with Pathogenic Potential. *Microbiol. Spectr.* 7. doi: 10.1128/microbiolspec.GPP3-0032-2018

Ehling-Schulz, M., Svensson, B., Guinebretiere, M. H., Lindback, T., Andersson, M., Schulz, A., et al. (2005). Emetic toxin formation of *Bacillus cereus* is restricted to a single evolutionary lineage of closely related strains. *Microbiology* 151, 183–197. doi: 10.1099/mic.0.27607-0

Elshaghabee, F. M. F., Rokana, N., Gulhane, R. D., Sharma, C., and Panwar, H. (2017). Bacillus as potential probiotics: status, concerns, and future perspectives. *Front. Microbiol.* 8:1490. doi: 10.3389/fmicb.2017.01490

Emms, D. M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16:157. doi: 10.1186/s13059-015-0721-2

Fagerlund, A., Lindback, T., and Granum, P. E. (2010). *Bacillus cereus* cytotoxins Hbl, Nhe and CytK are secreted via the Sec translocation pathway. *BMC Microbiol.* 10:304. doi: 10.1186/1471-2180-10-304

Fagerlund, A., Ween, O., Lund, T., Hardy, S. P., and Granum, P. E. (2004). Genetic and functional analysis of the cytK family of genes in *Bacillus cereus*. *Microbiology* 150, 2689–2697. doi: 10.1099/mic.0.26975-0

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Galili, T. (2015). dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* 31, 3718–3720. doi: 10.1093/bioinformatics/btv428

Gardner, S. N., and Hall, B. G. (2013). When whole-genome alignments just won't work: kSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes. *PLoS One* 8:e81760. doi: 10.1371/journal.pone.0081760

Gardner, S. N., Slezak, T., and Hall, B. G. (2015). kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics* 31, 2877–2878. doi: 10.1093/bioinformatics/btv271

Garnier, S. (2018). *Viridis: Default Color Maps from 'matplotlib. 0.5.1.* Available online at: https://CRAN.R-project.org/package=viridis (accessed April 11, 2020).

Gdoura-Ben Amor, M., Siala, M., Zayani, M., Grosset, N., Smaoui, S., Messadi-Akrout, F., et al. (2018). Isolation, identification, prevalence, and genetic diversity of *Bacillus cereus* group bacteria from different foodstuffs in Tunisia. *Front. Microbiol.* 9:447. doi: doi.org/10.3389/fmicb.2018.00447

Glasset, B., Herbin, S., Granier, S. A., Cavalie, L., Lafeuille, E., Guerin, C., et al. (2018). *Bacillus cereus*, a serious cause of nosocomial infections: epidemiologic and genetic survey. *PLoS One* 13:e0194346. doi: 10.1371/journal.pone.0194346

Glasset, B., Herbin, S., Guillier, L., Cadel-Six, S., Vignaud, M. L., Grout, J., et al. (2016). *Bacillus cereus*-induced food-borne outbreaks in France, 2007 to 2014: epidemiology and genetic characterisation. *Euro. Surveill.* 21:30413. doi: 10.2807/1560-7917.ES.2016.21.48.30413

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., and Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* 57, 81–91. doi: 10.1099/ijs.0.64483-0

Guinebretiere, M. H., Auger, S., Galleron, N., Contzen, M., De Sarrau, B., De Buyser, M. L., et al. (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. doi: 10.1099/ijs.0.030627-0

Guinebretiere, M.-H., Fagerlund, A., Granum, P. E., and Nguyen-The, C. (2006). Rapid discrimination of cytK-1 and cytK-2 genes in *Bacillus cereus* strains by a novel duplex PCR system. *FEMS Microbiol. Lett.* 259, 74–80. doi: 10.1111/j.1574-6968.2006.00247.x

Guinebretiere, M. H., Thompson, F. L., Sorokin, A., Normand, P., Dawyndt, P., Ehling-Schulz, M., et al. (2008). Ecological diversification in the *Bacillus cereus* Group. *Environ. Microbiol.* 10, 851–865. doi: 10.1111/j.1462-2920.2007.01495.x

Guinebretiere, M. H., Velge, P., Couvert, O., Carlin, F., Debuyser, M. L., and 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. doi: 10.1128/JCM.00921-10

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075. doi: 10.1093/bioinformatics/btt086

Hackathon, R., Bolker, B., Butler, M., Cowan, P., de Vienne, D., Eddelbuettel, D., Holder, M., et al. (2019). *phylobase: Base Package for Phylogenetic Structures and Comparative Data. R package version 0.8.10*. Available onlie at: https://CRAN.R-project.org/package=phylobase (accessed April 11, 2020).

Hanage, W. P., Fraser, C., and Spratt, B. G. (2005). Fuzzy species among recombinogenic bacteria. *BMC Biol.* 3:6. doi: 10.1186/1741-7007-3-6

Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* 35, 518–522. doi: 10.1093/molbev/msx281

Hoffmaster, A. R., Ravel, J., Rasko, D. A., Chapman, G. D., Chute, M. D., Marston, C. K., et al. (2004). Identification of anthrax toxin genes in a *Bacillus cereus* associated with an illness resembling inhalation anthrax. *Proc. Natl. Acad. Sci. U.S.A.* 101, 8449–8454. doi: 10.1073/pnas.0402414101

Huck, J. R., Hammond, B. H., Murphy, S. C., Woodcock, N. H., and Boor, K. J. (2007a). Tracking spore-forming bacterial contaminants in fluid milk-processing systems. *J. Dairy Sci.* 90, 4872–4883. doi: 10.3168/jds.2007-0196

Huck, J. R., Woodcock, N. H., Ralyea, R. D., and Boor, K. J. (2007b). Molecular subtyping and characterization of psychrotolerant endospore-forming bacteria in two New York state fluid milk processing systems. *J. Food Prot.* 70, 2354–2364. doi: 10.4315/0362-028X-70.10.2354

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., Von Mering, C., et al. (2017). Fast genome-wide functional annotation through Orthology assignment by egg NOG-Mapper. *Mol. Biol. Evol.* 34, 2115–2122. doi: 10.1093/molbev/msx148

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernandez-Plaza, A., Forslund, S. K., Cook, H., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47, D309–D314. doi: 10.1093/nar/gky1085

Huxley, J. (1943). Systematics and the origin of species : from the viewpoint of a zoologist. *Nature* 151, 347–348. doi: 10.1038/151347a0

Ivy, R. A., Ranieri, M. L., Martin, N. H., Den Bakker, H. C., Xavier, B. M., Wiedmann, M., et al. (2012). Identification and characterization of psychrotolerant sporeformers associated with fluid milk production and processing. *Appl. Environ. Microbiol.* 78, 1853–1864. doi: 10.1128/AEM.06536-11

Jain, C., Rodriguez, R. L., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* 9:5114. doi: 10.1038/s41467-018-07641-9

Jessberger, N., Krey, V. M., Rademacher, C., Bohm, M. E., Mohr, A. K., Ehling-Schulz, M., et al. (2015). From genome to toxicity: a combinatory approach highlights the complexity of enterotoxin production in *Bacillus cereus*. *Front. Microbiol.* 6:560. doi: 10.3389/fmicb.2015.00560

Jimenez, G., Urdiain, M., Cifuentes, A., Lopez-Lopez, A., Blanch, A. R., Tamames, J., et al. (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. doi: 10.1016/j.syapm.2013.04.008

Johler, S., Kalbhenn, E. M., Heini, N., Brodmann, P., Gautsch, S., Bagcioglu, M., et al. (2018). Enterotoxin production of *Bacillus thuringiensis* isolates from biopesticides, foods, and outbreaks. *Front Microbiol* 9:1915. doi: 10.3389/fmicb.2018.01915

Jolley, K. A., Bray, J. E., and Maiden, M. C. J. (2018). Open-access bacterial population genomics: BIGSdb software, the PubMLST.org *website and their applications*. *Wellcome Open Res.* 3:124. doi: 10.12688/wellcomeopenres.14826.1

Jolley, K. A., and Maiden, M. C. (2010). BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 11:595. doi: 10.1186/1471-2105-11-595

Jombart, T., Kendall, M., Almagro-Garcia, J., and Colijn, C. (2017). treespace: Statistical exploration of landscapes of phylogenetic trees. *Mol. Ecol. Resour.* 17, 1385–1392. doi: 10.1111/1755-0998.12676

Jouzani, G. S., Valijanian, E., and Sharafi, R. (2017). *Bacillus thuringiensis*: a successful insecticide with new environmental features and tidings. *Appl. Microbiol. Biotechnol.* 101, 2691–2711. doi: 10.1007/s00253-017-8175-y

Jung, M. Y., Kim, J. S., Paek, W. K., Lim, J., Lee, H., Kim, P. I., et al. (2011). *Bacillus manliponensis* sp. nov., a new member of the *Bacillus cereus* group isolated from foreshore tidal flat sediment. *J. Microbiol* .49, 1027–1032. doi: 10.1007/s12275-011-1049-6

Jung, M. Y., Paek, W. K., Park, I. S., Han, J. R., Sin, Y., Paek, J., et al. (2010). *Bacillus gaemokensis* sp. nov., isolated from foreshore tidal flat sediment from the Yellow Sea. *J. Microbiol.* 48, 867–871. doi: 10.1007/s12275-010-0148-0

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Katz, L. S., Griswold, T., Williams-Newkirk, A. J., Wagner, D., Petkau, A., Sieffert, C., et al. (2017). A comparative analysis of the Lyve-SET phylogenomics pipeline for genomic epidemiology of foodborne pathogens. *Front. Microbiol.* 8:375. doi: 10.3389/fmicb.2017.00375

Kendall, M., and Colijn, C. (2015). A tree metric using structure and length to capture distinct phylogenetic signals. *arXiv* [Preprint].

Kendall, M., and Colijn, C. (2016). Mapping phylogenetic trees to reveal distinct patterns ofe. *Mol. Biol. Evol.* 33, 2735–2743. doi: 10.1093/molbev/msw124

Kindle, P., Etter, D., Stephan, R., and Johler, S. (2019). Population structure and toxin gene profiles of *Bacillus cereus* sensu lato isolated from flour products. *FEMS Microbiol. Lett.* 366:fnz240. doi: 10.1093/femsle/fnz240

Klee, S. R., Brzuszkiewicz, E. B., Nattermann, H., Bruggemann, H., Dupke, S., Wollherr, A., et al. (2010). The genome of a Bacillus isolate causing anthrax in chimpanzees combines chromosomal properties of B. *cereus with B. anthracis* virulence plasmids. *PLoS One* 5:e10986. doi: 10.1371/journal.pone.0010986

Kovac, J., Miller, R. A., Carroll, L. M., Kent, D. J., Jian, J., Beno, S. M., et al. (2016). Production of hemolysin BL by *Bacillus cereus* group isolates of dairy origin is associated with whole-genome phylogenetic clade. *BMC Genomics* 17:581. doi: 10.1186/s12864-016-2883-z

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika* 29, 115–129. doi: 10.1007/BF02289694

Lazarte, J. N., Lopez, R. P., Ghiringhelli, P. D., and Beron, C. M. (2018). Bacillus wiedmannii biovar thuringiensis: a specialized mosquitocidal pathogen with plasmids from diverse origins. *Genome Biol. Evol.* 10, 2823–2833. doi: 10.1093/gbe/evy211

Lechner, S., Mayr, R., Francis, K. P., Pruss, B. M., Kaplan, T., Wiessner-Gunkel, E., et al. (1998). Bacillus weihenstephanensis sp. nov. is a new psychrotolerant species of the *Bacillus cereus* group. *Int. J. Syst. Bacteriol.* 48(Pt. 4), 1373–1382. doi: 10.1099/00207713-48-4-1373

Lee, I., Ouk Kim, Y., Park, S. C., and Chun, J. (2016). OrthoANI: an improved algorithm and software for calculating average nucleotide identity. *Int. J. Syst. Evol. Microbiol.* 66, 1100–1103. doi: 10.1099/ijsem.0.000760

Leendertz, F. H., Ellerbrok, H., Boesch, C., Couacy-Hymann, E., Matz-Rensing, K., Hakenbeck, R., et al. (2004). Anthrax kills wild chimpanzees in a tropical rainforest. *Nature* 430, 451–452. doi: 10.1038/nature02722

Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* 50, 913–925. doi: 10.1080/106351501753462876

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Liu, B., Liu, G. H., Hu, G. P., Sengonca, C., Lin, N. Q., Tang, J. Y., et al. (2014). *Bacillus bingmayongensis* sp. nov., isolated from the pit soil of Emperor Qin's Terra-cotta warriors in China. *Antonie Van Leeuwenhoek* 105, 501–510. doi: 10.1007/s10482-013-0102-3

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

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2019). cluster: Cluster Analysis Basics and Extensions. *R. Package Version* 1, 56.

Marston, C. K., Ibrahim, H., Lee, P., Churchwell, G., Gumke, M., Stanek, D., et al. (2016). Anthrax toxin-expressing *Bacillus cereus* isolated from an Anthrax-like eschar. *PLoS One* 11:e0156987. doi: 10.1371/journal.pone.0156987

Messelhäußer, U., and Ehling-Schulz, M. (2018). *Bacillus cereus*—a multifaceted opportunistic pathogen. *Curr. Clin. Microbiol. Rep.* 5, 120–125. doi: 10.1007/s40588-018-0095-9

Miller, R. A., Beno, S. M., Kent, D. J., Carroll, L. M., Martin, N. H., Boor, K. J., et al. (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. doi: 10.1099/ijsem.0.001421

Miller, R. A., Jian, J., Beno, S. M., Wiedmann, M., and Kovac, J. (2018). Intraclade variability in toxin production and cytotoxicity of *Bacillus cereus* group type strains and dairy-associated isolates. *Appl. Environ. Microbiol.* 84:e02479–17. doi: 10.1128/AEM.02479-17

Minh, B. Q., Nguyen, M. A., and Von Haeseler, A. (2013). Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* 30, 1188–1195. doi: 10.1093/molbev/mst024

Moayeri, M., Leppla, S. H., Vrentas, C., Pomerantsev, A. P., and Liu, S. (2015). Anthrax Pathogenesis. *Annu. Rev. Microbiol.* 69, 185–208. doi: 10.1146/annurev-micro-091014-104523

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Oh, S. Y., Budzik, J. M., Garufi, G., and Schneewind, O. (2011). Two capsular polysaccharides enable *Bacillus cereus* G9241 to cause anthrax-like disease. *Mol. Microbiol.* 80, 455–470. doi: 10.1111/j.1365-2958.2011.07582.x

Okinaka, R. T., Cloud, K., Hampton, O., Hoffmaster, A. R., Hill, K. K., Keim, P., et al. (1999). Sequence and organization of pXO1, the large Bacillus anthracis plasmid harboring the anthrax toxin genes. *J. Bacteriol.* 181, 6509–6515. doi: 10.1128/JB.181.20.6509-6515.1999

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., Mcglinn, D., et al. (2019). *vegan: Community Ecology Package. R package version* 2.5-6. https://CRAN.R-project.org/package=vegan.

Ozdemir, F., and Arslan, S. (2019). Molecular characterization and toxin profiles of Bacillus spp. isolated from retail fish and ground beef. *J. Food Sci.* 84, 548–556. doi: 10.1111/1750-3841.14445

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289–290. doi: 10.1093/bioinformatics/btg412

Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528. doi: 10.1093/bioinformatics/bty633

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–1055. doi: 10.1101/gr.186072.114

Pilo, P., and Frey, J. (2011). Bacillus anthracis: molecular taxonomy, population genetics, phylogeny and patho-evolution. *Infect. Genet. Evol.* 11, 1218–1224. doi: 10.1016/j.meegid.2011.05.013

Pilo, P., and Frey, J. (2018). Pathogenicity, population genetics and dissemination of Bacillus anthracis. *Infect. Genet. Evol.* 64, 115–125. doi: 10.1016/j.meegid.2018.06.024

Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G., and Toth, I. K. (2016). Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal. Methods* 8, 12–24. doi: 10.1039/C5AY02550H

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35, D61–D65. doi: 10.1093/nar/gkl842

Rasko, D. A., Altherr, M. R., Han, C. S., and Ravel, J. (2005). Genomics of the *Bacillus cereus* group of organisms. *FEMS Microbiol. Rev.* 29, 303–329. doi: 10.1016/j.femsre.2004.12.005

R Core Team, (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* 3, 217–223. doi: 10.1111/j.2041-210X.2011.00169.x

Richter, M., and Rossello-Mora, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. U.S.A.* 106, 19126–19131. doi: 10.1073/pnas.0906412106

Richter, M., Rossello-Mora, R., Oliver Glockner, F., and Peplies, J. (2016). JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics* 32, 929–931. doi: 10.1093/bioinformatics/btv681

Rigaux, C., Ancelet, S., Carlin, F., Nguyen-The, C., and Albert, I. (2013). Inferring an augmented Bayesian network to confront a complex quantitative microbial risk assessment model with durability studies: application to *Bacillus cereus* on a courgette puree production chain. *Risk Anal.* 33, 877–892. doi: 10.1111/j.1539-6924.2012.01888.x

Riol, C. D., Dietrich, R., Martlbauer, E., and Jessberger, N. (2018). Consumed Foodstuffs Have a Crucial Impact on the Toxic Activity of Enteropathogenic *Bacillus cereus*. *Front. Microbiol.* 9:1946. doi: 10.3389/fmicb.2018.01946

Rocha, E. P. C. (2018). Neutral theory, microbial practice: challenges in bacterial population genetics. *Mol. Biol. Evol.* 35, 1338–1347. doi: 10.1093/molbev/msy078

Rodriguez, R. L., Gunturu, S., Harvey, W. T., Rossello-Mora, R., Tiedje, J. M., Cole, J. R., et al. (2018). The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. *Nucleic Acids Res.* 46, W282–W288. doi: 10.1093/nar/gky467

Romero-Alvarez, D., Peterson, A. T., Salzer, J. S., Pittiglio, C., Shadomy, S., Traxler, R., et al. (2020). Potential distributions of Bacillus anthracis and *Bacillus cereus* biovar anthracis causing anthrax in Africa. *PLoS Negl. Trop. Dis.* 14:e0008131.

Rosvall, M., Axelsson, D., and Bergstrom, C. T. (2009). The map equation. *Eur. Phys. J. Spec. Top.* 178, 13–23. doi: 10.1140/epjst/e2010-01179-1

Rouzeau-Szynalski, K., Stollewerk, K., Messelhausser, U., and Ehling-Schulz, M. (2020). Why be serious about emetic *Bacillus cereus*: cereulide production and industrial challenges. *Food Microbiol.* 85:103279. doi: 10.1016/j.fm.2019.103279

Schoeni, J. L., and Wong, A. C. (2005). *Bacillus cereus* food poisoning and its toxins. *J. Food Prot.* 68, 636–648. doi: 10.4315/0362-028X-68.3.636

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Senesi, S., and Ghelardi, E. (2010). Production, secretion and biological activity of *Bacillus cereus* enterotoxins. *Toxins* 2, 1690–1703. doi: 10.3390/toxins2071690

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

Stevens, M. J. A., Tasara, T., Klumpp, J., Stephan, R., Ehling-Schulz, M., and 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* 9, 1984. doi: 10.1038/s41598-018-36254-x

Tallent, S. M., Knolhoff, A., Rhodehamel, E. J., Harmon, S. M., and Bennett, R. W. (2019). *Chapter 14: Bacillus cereus in Bacteriological Analytical Manual (BAM)*. Silver Spring, ML: Food and Drug Administration.

Tallent, S. M., Kotewicz, K. M., Strain, E. A., and Bennett, R. W. (2012). Efficient isolation and identification of *Bacillus cereus* group. *J. AOAC Int.* 95, 446–451. doi: 10.5740/jaoacint.11-251

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci.* 17, 57–86.

The Gene Ontology Consortium. (2018). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* 47, D330–D338. doi: 10.1093/nar/gky1055

Thorsen, L., Kando, C. K., Sawadogo, H., Larsen, N., Diawara, B., Ouedraogo, G. A., et al. (2015). Characteristics and phylogeny of *Bacillus cereus* strains isolated from Maari, a traditional West African food condiment. *Int. J. Food Microbiol.* 196, 70–78. doi: 10.1016/j.ijfoodmicro.2014.11.026

Tonkin-Hill, G., Lees, J. A., Bentley, S. D., Frost, S. D. W., and Corander, J. (2018). RhierBAPS: An R implementation of the population clustering algorithm hierBAPS. *Wellcome Open Res* 3:93. doi: 10.12688/wellcomeopenres.14694.1

Ulrich, S., Gottschalk, C., Dietrich, R., Martlbauer, E., and Gareis, M. (2019). Identification of cereulide producing *Bacillus cereus* by MALDI-TOF MS. *Food Microbiol.* 82, 75–81. doi: 10.1016/j.fm.2019.01.012

Webb, M. D., Barker, G. C., Goodburn, K. E., and Peck, M. W. (2019). Risk presented to minimally processed chilled foods by psychrotrophic *Bacillus cereus*. *Trends Food Sci. Technol.* 93, 94–105. doi: 10.1016/j.tifs.2019.08.024

Wickham, H. (2007). Reshaping Data with the reshape Package. *J. Statist. Softw.* 21, 1–20. doi: 10.18637/jss.v021.i12

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag. doi: 10.1007/978-3-319-24277-4

Wickham, H., and Bryan, J. (2019). *Readxl: Read Excel Files. R package version 1.3.1*. available online at: https://CRAN.R-project.org/package=readxl (accessed April 11, 2020).

Wickham, H., François, R., Henry, L., and Müller, K. (2020). *dplyr: A Grammar of Data Manipulation. R package version 0.8.5*. Available online at: https://CRAN.R-project.org/package=dplyr (accessed April 11, 2020).

Wilson, M. K., Vergis, J. M., Alem, F., Palmer, J. R., Keane-Myers, A. M., Brahmbhatt, T. N., et al. (2011). *Bacillus cereus* G9241 makes anthrax toxin and capsule like highly virulent *B. anthracis* Ames but behaves like attenuated toxigenic nonencapsulated *B. anthracis* Sterne in rabbits and mice. *Infect. Immun.* 79, 3012–3019. doi: 10.1128/IAI.00205-11

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39, 306–314. doi: 10.1007/BF00160154

Yoon, S. H., Ha, S. M., Lim, J., Kwon, S., and Chun, J. (2017). A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie Van Leeuwenhoek* 110, 1281–1286. doi: 10.1007/s10482-017-0844-4

Yu, G., Lam, T. T., Zhu, H., and Guan, Y. (2018). Two methods for mapping and visualizing associated data on phylogeny using Ggtree. *Mol. Biol. Evol.* 35, 3041–3043. doi: 10.1093/molbev/msy194

Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T.-Y. (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* 8, 28–36. doi: 10.1111/2041-210X.12628

Zhao, C., and Wang, Z. (2018). GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms. *Sci. Rep.* 8:15107. doi: 10.1038/s41598-018-33219-y

Zhao, S., Chen, J., Fei, P., Feng, H., Wang, Y., Ali, M. A., et al. (2020). Prevalence, molecular characterization, and antibiotic susceptibility of *Bacillus cereus* isolated from dairy products in China. *J. Dairy Sci.* 103, 3994–4001. doi: 10.3168/jds.2019-17541

Zhuang, K., Li, H., Zhang, Z., Wu, S., Zhang, Y., Fox, E. M., et al. (2019). Typing and evaluating heat resistance of *Bacillus cereus* sensu stricto isolated from the processing environment of powdered infant formula. *J. Dairy Sci.* 102, 7781–7793. doi: 10.3168/jds.2019-16392

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.