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], Jasna Kovac[3]

[1]Structural and Computational Biology Unit, EMBL, Heidelberg, Germany

[2]Department of Food Science, Cornell University, Ithaca, NY, USA

[3]Department of Food Science, The Pennsylvania State University, University Park, PA, USA

*Correspondence: Jasna Kovac, jzk303@psu.edu

Keywords: *Bacillus anthracis*, *Bacillus cereus*, *Bacillus cereus* group, *Bacillus thuringiensis*, foodborne illness, taxonomy

Abstract

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), a technology with limited accessibility. 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) *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 congruent with the proposed eight-genomospecies taxonomy and resolves incongruencies observed in the historical seven-group framework among isolates assigned to *panC* Groups II, III, and VI. 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

39 units within this proposed framework with little-to-no user intervention or domain-specific
40 knowledge of *B. cereus s.l.* taxonomy. We additionally outline a novel method for assigning *B.*
41 *cereus s.l.* genomes to pseudo-gene flow units within proposed genomospecies. The results
42 presented here highlight the backwards-compatibility and accessibility of the proposed
43 taxonomic framework and illustrate that WGS is not a necessity for microbiologists who want to
44 use the proposed taxonomy effectively.
45

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; Moayeri et al., 2015; Pilo and Frey, 2018). 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.), 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., 2020). 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*) 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 (i.e., subspecies *anthracis*, which is used to refer to the classic non-motile, non-hemolytic lineage referred to as "*B. anthracis*", and subspecies *cereus*, which is used to refer to

panC Group III 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 used 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., 2020). 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., 2020), 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 backwards-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.

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., 2020) 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 a genomospecies, subspecies (if applicable), and biovar(s) (if applicable), using a recently proposed taxonomy (Carroll et al., 2020). 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., 2020) or an effective or proposed *B. cereus s.l.* genomospecies (i.e., "*B. bingmayongensis*", "*B. clarus*", *"B. gaemokensis*", or *"B. manliponensis*"), were excluded (Jung et al., 2010; Jung et al., 2011; Liu et al., 2014; Acevedo et al., 2019), yielding a set of 1,741 high-quality *B. cereus s.l.* genomes. 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., 2020).

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., thresholds of 92.5, 94, 95, and 96 ANI) (Guinebretiere et al., 2013; Jimenez et al., 2013; Miller et al., 2016; Liu et al., 2017; Carroll et al., 2020) as described previously (Carroll et al., 2020), using the bactaxR package (Carroll et al., 2020) in R v. 3.6.1 (R Core Team, 2019) and 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; Yu et al., 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), and 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. and BTyper v. 2.3.3 (Guinebretiere et al., 2008; Guinebretiere et al., 2010; 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., 2020). 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., 2020). 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. (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) ≥ 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 ≥ 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; Huerta-Cepas et al., 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 (http://dna.cs.miami.edu/GOGO/; accessed May 30, 2020) 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 were 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 (Figure 1B and Figure 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 (Figure 1B and Figure 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., 2020), 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. 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 improve *in silico* virulence factor detection in *B. cereus s.l.* genomes, the BTyper3 v. 3.1.0 virulence factor database was constructed to include amino acid sequences of the following virulence factors: (i) anthrax toxin genes *cya*, *lef*, and *pagA* (the same sequences used for assignment of biovar Anthracis in all previous versions of BTyper3); (ii) cereulide synthetase

319 genes *cesABCD* (the same sequences used for assignment of biovar Emeticus in all previous
320 versions of BTyper3); (iii) non-hemolytic enterotoxin (Nhe) genes *nheABC* (used in the original
321 BTyper v. 2.3.3 and earlier); (iv) hemolysin BL (Hbl) genes *hblABCD* (used in the original
322 BTyper v. 2.3.3 and earlier); (v) cytotoxin K (CytK) variant 1 and 2 (*cytK-1* and *cytK-2*,
323 respectively; used in the original BTyper v. 2.3.3 and earlier); (vi) sphingomyelinase Sph gene
324 *sph* (used in the original BTyper v. 2.3.0-2.3.3); (vii) anthrax capsule biosynthesis (Cap) genes
325 *capABCDE* (used in the original BTyper v. 2.3.3 and earlier); (viii) hyaluronic acid capsule
326 (Has) genes *has ABC* (*hasA* was included in the original BTyper v. 2.3.3 and earlier, and *hasBC*
327 were added here) (Oh et al., 2011); (ix) exo-polysaccharide capsule (Bps) genes
328 *bpsXABCDEFGH* (used in the original BTyper v 2.0.1-2.3.3).

329 To provide updated boundaries for virulence factor detection based on a larger set of
330 genomes that span *B. cereus s.l.*, BTyper3 v. 3.1.0 was used to identify all virulence factors listed
331 above in the complete set of 1,741 high-quality genomes (see section "Acquisition of *Bacillus*
332 *cereus s.l.* genomes" above), using a maximum BLAST E-value threshold of 1E-5 (Carroll et al.,
333 2017; Carroll et al., 2020), but with minimum amino acid identity and coverage thresholds of 0%
334 each. Plots of virulence factors detected within all genomes at various amino acid identity and
335 coverage thresholds were constructed using ggplot2 in R (Figure 3 and Supplementary Table
336 S3). Based on these plots, amino acid identity and coverage thresholds of 70% and 80%,
337 respectively, were implemented as the default thresholds for virulence factor detection in
338 BTyper3 v. 3.1.0 (Figure 3). Additionally, to reduce the risk of users mis-interpreting spurious
339 hits that do not correlate with a virulence phenotype, BTyper3 v. 3.1.0 reports virulence factors
340 at an operon/cluster level; for example, if only *cesC* is detected in a genome, BTyper3 reports
341 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% 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 and 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 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

388 strain genomes and genomospecies thresholds of 94-96 ANI were assigned to the *B. mycoides*
389 genomospecies cluster (Table 1 and Supplementary Table S4). Additionally, all genomes that
390 shared 94-96 ANI with the *B. cereus s.s.* str. ATCC 14579 and/or *B. thuringiensis* serovar
391 berliner str. ATCC 10792 type strain genomes belonged to the *B. cereus s.s.* genomospecies
392 cluster (Table 1 and Supplementary Table S4). However, it should be noted that the "*B. cereus*"
393 and "*B. thuringiensis*" species as historically defined are polyphyletic, and other strains often
394 referred to as "*B. cereus*" or "*B. thuringiensis*" belong to other genomospecies clusters; emetic
395 reference strain "*B. cereus*" str. AH187, for example, belongs to *B. mosaicus* and not *B. cereus*
396 *s.s.* (Carroll et al., 2019; Carroll et al., 2020).
397 **STs assigned using seven-gene MLST can be used for *B. cereus s.l.* genomospecies**
398 **assignment**. All STs assigned using BTyper3 and PubMLST's seven-gene MLST scheme *for B.*
399 *cereus s.l.* (Jolley and Maiden, 2010; Jolley et al., 2018) were contained within a single proposed
400 *B. cereus s.l.* genomospecies, and no STs were split across multiple genomospecies
401 (Supplementary Tables S1 and S5). As such, a comprehensive list of ST/genomospecies
402 associations for all NCBI RefSeq *B. cereus s.l.* genomes is available (n = 2,231; RefSeq accessed
403 November 19, 2018, PubMLST *B. cereus* database accessed April 26, 2020; Supplementary
404 Tables S1 and S5). However, it is essential to note that the *B. cereus s.l.* phylogeny constructed
405 using the sequences of these seven alleles alone (i.e., the MLST phylogeny) did not mirror the
406 WGS-based *B. cereus s.l.* phylogeny perfectly. Regardless of the ANI threshold used (i.e., 92.5,
407 94, 95, or 96 ANI), the *B. cereus s.l.* MLST phylogeny yielded polyphyletic genomospecies
408 clusters (Figure 5 and Supplementary Figures S1-S5), although genomospecies clusters formed
409 at 92.5 ANI reduced the proportion of polyphyletic genomospecies within the MLST phylogeny.
410 One of eight genomospecies (12.5%) defined at 92.5 ANI were polyphyletic based on the MLST

411 tree, compared to 2/11 (18.2%), 3/21 (14.3%), and 4/30 (13.3%) polyphyletic genomospecies
412 defined at 94, 95, and 96 ANI respectively (Figure 5 and Supplementary Figures S1-S5).
413 **An adjusted, eight-group *panC* framework remains largely congruent with proposed *B.***
414 ***cereus s.l.* genomospecies definitions.** Another popular typing method used to assign *B. cereus*
415 *s.l.* isolates to major phylogenetic groups relies on the sequence of *panC* (Guinebretiere et al.,
416 2008; Guinebretiere et al., 2010). However, the seven-group *panC* framework had to be adjusted
417 to accommodate the growing amount of *B. cereus s.l.* genomic diversity provided by WGS, as
418 *panC* sequences assigned to Groups II, III, and VI using the seven-group typing scheme
419 implemented in the original BTyper were polyphyletic (Figure 4A).
420 The adjusted, eight-group *panC* framework constructed here (Figure 4B) and
421 implemented in BTyper3 v. 3.1.0 resolved all polyphyletic *panC* group assignments (Figure 4).
422 *panC* group assignments using the adjusted, eight-group framework described here, as well as
423 those obtained using the original seven-group framework implemented in BTyper v. 2.3.3, are
424 available for 2,229 *B. cereus s.l.* genomes (Table 1 and Supplementary Tables S1 and S6). Note
425 that group assignments using the seven-group framework implemented in the web-tool published
426 by Guinebretiere, et al. (Guinebretiere et al., 2010) are not available, as the database is not
427 publicly available, and the web-based method is not scalable.
428 However, even with an improved eight-group framework for *panC* group assignment, the
429 *B. cereus s.l. panC* phylogeny yielded polyphyletic genomospecies, regardless of the ANI-based
430 threshold used to define genomospecies. For seven of the eight *B. cereus s.l.* genomospecies
431 defined at 92.5 ANI (with the exclusion of effective and proposed putative species), the *panC*
432 locus produced a monophyletic clade for each genomospecies (Figures 4 and 5 and
433 Supplementary Figures S6 and S7). However, based on the sequence of *panC*, the *B. mosaicus*

434 genomospecies was polyphyletic, with the *panC* sequence of *B. luti* forming a separate lineage
435 within the *B. mosaicus panC* clade (Figure 5 and Supplementary Figures S6 and S7). Similarly,
436 genomospecies defined at 94, 95, and 96 ANI produced even greater proportions of polyphyletic
437 *panC* clusters, with 5/11 (45.5%), 8 or 9/21 (38.1 or 42.9%, depending on the phylogeny rooting
438 method), and 8/30 (26.7%) genomospecies polyphyletic via *panC*, respectively (Supplementary
439 Figures S6-S15).
440 ***rpoB* provides lower resolution than *panC* for single-locus sequence typing of *B. cereus s.l.***
441 **isolates.** Another popular single-locus sequence typing method for characterizing spore-forming
442 bacteria, including *B. cereus s.l.* isolates, relies on sequencing *rpoB*, which encodes the beta
443 subunit of RNA polymerase (Huck et al., 2007b; Ivy et al., 2012). Among publicly available *B.*
444 *cereus s.l.* isolate genomes, ATs assigned using the Cornell University Food Safety Laboratory
445 and Milk Quality Improvement Program's (CUFSL/MQIP) *rpoB* allelic typing database (Carroll
446 et al., 2017), much like STs assigned using PubMLST's seven-gene scheme (described above),
447 were each confined to a single genomospecies at 92.5 ANI, with no AT split across
448 genomospecies (Supplementary Tables S1 and S7). However, fewer than 2/3 of all *B. cereus s.l.*
449 genomes possessed a *rpoB* allele that matched a member of the database exactly (i.e., with 100%
450 nucleotide identity and coverage; 1,425/2,231 genomes, or 63.9%). Additionally, the *B. cereus*
451 *s.l. rpoB* phylogeny showcased numerous polyphyletic genomospecies, regardless of the ANI
452 threshold at which genomospecies were defined (3/8 [37.5%], 3 or 4/11 [27.2 or 36.4%,
453 depending on the phylogeny rooting method], 6 or 9/21 [28.6 or 42.9%, depending on the
454 phylogeny rooting method], and 8/30 [26.7%] polyphyletic *rpoB* clades among genomospecies
455 defined at 92.5, 94, 95, and 96 ANI, respectively; Figure 5 and Supplementary Figures S16-S23).

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 ≥ 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).

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; Ehling-Schulz et al., 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)

479 *panC* Group III and VI, respectively, as described previously (Guinebretiere et al., 2008;
480 Guinebretiere et al., 2010; Carroll et al., 2017; Carroll and Wiedmann, 2020; Carroll et al., 2020)
481 and regardless of whether the legacy seven-group or adjusted eight-group *panC* typing schemes
482 were used (Figure 7 and Supplementary Table S1).

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

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; Stenfors Arnesen et al., 2008; Guinebretiere et al., 2010; 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 94.7-98.9 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

525 the query genome shares an ANI value with one of the gene flow unit medoid genomes that is
526 greater than or equal to the previously observed ANI boundary for the gene flow unit, it is
527 assigned to that particular pseudo-gene flow unit (Figures 1 and 2 and Table 2). This pseudo-
528 gene flow unit assignment method was applied to all 2,231 *B. cereus s.l.* genomes (Table 2 and
529 Supplementary Table S1), and was found to yield pseudo-gene flow units that were each
530 encompassed within a single genomospecies and *panC* group (using the adjusted eight-group
531 *panC* scheme developed here), with no pseudo-gene flow units split across multiple
532 genomospecies/*panC* groups (Table 2). PopCOGenT identified multiple gene flow units among
533 the *B. cereus s.s.*, *B. mosaicus, B. mycoides,* and *B. toyonensis* genomospecies delineated at 92.5
534 ANI (n = 4, 16, 7, and 2 main clusters, respectively; Figure 8 and Supplementary Figures S25-
535 S32).
536
537 **Discussion**
538 **The proposed *B. cereus s.l.* taxonomy is backwards-compatible with *B. cereus s.l.* species**
539 **defined using historical ANI-based species thresholds.** ANI-based methods have been used to
540 define 12 *B. cereus s.l.* species prior to 2020: *B. cytotoxicus* and *B. toyonensis*, each proposed as
541 novel species in 2013 (Guinebretiere et al., 2013; Jimenez et al., 2013), *B. wiedmannii* (proposed
542 as a novel species in 2016) (Miller et al., 2016), and nine species (*B. albus, B. luti, B. mobilis, B.*
543 *nitratireducens, B. pacificus, B. paranthracis*, *B. paramycoides, B. proteolyticus,* and *B.*
544 *tropicus*) proposed in 2017 (Liu et al., 2017). However, the lack of a standardized ANI-based
545 genomospecies threshold for defining *B. cereus s.l.* genomospecies has led to confusion
546 regarding how *B. cereus s.l.* species should be delineated. *B. toyonensis* and the nine species
547 proposed in 2017, for example, were defined using genomospecies thresholds of 94 and 96 ANI,

548 respectively (Jimenez et al., 2013; Liu et al., 2017). The descriptions of *B. cytotoxicus* and *B.*
549 *wiedmannii* as novel species each explicitly state that a 95 ANI threshold was used
550 (Guinebretiere et al., 2013; Miller et al., 2016); however, the *B. wiedmannii* type strain genome
551 shared a much higher degree of similarity with the type strain genome of its neighboring species
552 than did *B. cytotoxicus* (Miller et al., 2016). As such, choice of ANI-based genomospecies
553 threshold can affect which *B. cereus s.l.* strains belong to which genomospecies, and may even
554 produce overlapping genomospecies in which a genome can belong to more than one
555 genomospecies (Carroll et al., 2020).

556 The proposed *B. cereus s.l.* taxonomy (Carroll et al., 2020) provides a standardized
557 genomospecies threshold of 92.5 ANI, which has been shown to yield non-overlapping,
558 monophyletic *B. cereus s.l.* genomospecies. However, the practice of assigning *B. cereus s.l.*
559 isolates to genomospecies using species type strain genomes and historical species thresholds
560 (i.e., 94-96 ANI) has been important for whole-genome characterization for *B. cereus s.l.* strains,
561 including those responsible for illnesses and/or outbreaks (Lazarte et al., 2018; Bukharin et al.,
562 2019; Carroll et al., 2019). Here, we show that all 18 published *B. cereus s.l.* genomospecies
563 defined prior to 2020 are safely integrated into the proposed *B. cereus s.l.* taxonomy without
564 polyphyly, regardless of whether a 94, 95, or 96 ANI genomospecies threshold was used to
565 delineate species relative to type strain genomes.

566 **Single- and multi-locus sequence typing methods can be used to assign *B. cereus s.l.* isolates**
567 **to species within the proposed taxonomy.** Single- and multi-locus sequence typing approaches
568 have been—and continue to be—important methods for classifying *B. cereus s.l.* isolates into
569 phylogenetic units. They have been used to characterize *B. cereus s.l.* strains associated with
570 illnesses and outbreaks (Cardazzo et al., 2008; Glasset et al., 2016; Akamatsu et al., 2019;

571 Carroll et al., 2019), strains isolated from food and food processing environments (Huck et al.,
572 2007a; Thorsen et al., 2015; Kindle et al., 2019; Ozdemir and Arslan, 2019; Zhuang et al., 2019;
573 Zhao et al., 2020), and strains with industrial applications (e.g., biopesticide strains) (Johler et
574 al., 2018). Additionally, STs and ATs assigned using these approaches have been used to
575 construct frameworks for predicting the risk that a particular *B. cereus s.l.* strain poses to food
576 safety, public health, and food spoilage (Guinebretiere et al., 2010; Rigaux et al., 2013; Buehler
577 et al., 2018; Miller et al., 2018; Webb et al., 2019). It is thus important to ensure that the
578 proposed standardized taxonomy for *B. cereus s.l.* remains congruent with widely used sequence
579 typing approaches.

580 Here, we assessed the congruency of three popular single- and multi-locus sequence
581 typing schemes for *B. cereus s.l.* with proposed genomospecies definitions: (i) the PubMLST
582 seven-gene MLST scheme for *B. cereus s.l.* (Jolley and Maiden, 2010; Jolley et al., 2018), (ii)
583 the seven-group *panC* typing scheme developed by Guinebretierere, et al. (Guinebretiere et al.,
584 2008; Guinebretiere et al., 2010) as implemented in the original BTyper (Carroll et al., 2017),
585 and (iii) the CUFSL/MQIP *rpoB* allelic typing scheme used for characterizing spore-forming
586 bacteria, including members of *B. cereus s.l.* (Durak et al., 2006; Huck et al., 2007a; Ivy et al.,
587 2012; Buehler et al., 2018). STs and ATs assigned using MLST and *rpoB* allelic typing,
588 respectively, as well as six of eight *panC* groups assigned using the adjusted eight-group
589 framework developed here, were each contained within a single genomospecies at 92.5 ANI.
590 Thus, past studies employing these methods can be easily interpreted within the proposed
591 taxonomic framework for the group.

592 MLST, *panC* group assignment, and *rpoB* allelic typing will likely remain extremely
593 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. (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 backwards-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

640 "pseudo-gene flow unit" method in BTyper3 v. 3.1.0 that can be used to assign a user's genome
641 of interest to a pseudo-gene flow unit using the set of 33 PopCOgenT gene flow unit medoid
642 genomes, the pairwise ANI values calculated within PopCOgenT gene flow units, and FastANI.
643 However, it is essential to note the limitations of the pseudo-gene flow unit assignment method
644 implemented in BTyper3. First and foremost, ANI and the methods employed by PopCOgenT
645 are fundamentally and conceptually different; the pseudo-gene flow unit assignment method
646 described here does not infer recent gene flow, nor does it use PopCOgenT or any of its metrics.
647 Thus, the pseudo-gene flow unit assignment method cannot be used to construct true gene flow
648 units for *B. cereus s.l.* Secondly, to increase the speed of PopCOGenT, we reduced *B. cereus s.l.*
649 to a set of 313 representative genomes that encompassed the diversity of the species complex;
650 genomes that shared ≥ 99 ANI with one or more genomes in this representative set were omitted
651 (i.e., 1,428 of 1,741 high-quality genomes were omitted; 82.0%). Consequently, gene flow
652 among most closely related lineages that shared ≥ 99 ANI with each other was not assessed, as it
653 was thereby assumed that highly similar genomes that shared ≥ 99 ANI with each other belonged
654 to the same PopCOGenT "main cluster" (i.e., "species"). It is possible that the inclusion of these
655 highly similar genomes would have resulted in the discovery of additional gene flow units, or
656 perhaps changes in existing ones, and future studies are needed to assess and refine this.
657 However, the pseudo-gene flow unit assignment approach described here allows researchers to
658 rapidly identify the most similar medoid genome of true gene flow units identified within *B.*
659 *cereus s.l.* Results should not be interpreted as an assessment of recent gene flow, but rather as a
660 higher-resolution phylogenomic clade assignment, similar to how one might use MLST for
661 delineation of lineages within species. We anticipate that our rapid method will be valuable to
662 researchers who desire greater resolution than what is provided at the genomospecies level,

663 particularly when querying diverse *B. cereus s.l.* genomospecies that comprise multiple major
664 clades (e.g., *B. mosaicus*, *B. mycoides, B. cereus s.s.*).
665
666 **Author Contributions**
667 LMC performed all computational analyses. LMC, RAC, and JK designed the study and co-
668 wrote the manuscript.
669
670 **Funding**
671 This work was supported by USDA NIFA Hatch Appropriations under project no. PEN04646
672 and accession no. 1015787, and the USDA NIFA grant GRANT12686965.
673
674 **Conflict of Interest Statement**
675 The authors declare that the research was conducted in the absence of any personal, professional,
676 or financial relationships that could potentially be construed as a conflict of interest.
677
678
679
680
681
682
683
684
685

686 **References**

687

688 Acevedo, M.M., Carroll, L.M., Mukherjee, M., Mills, E., Xiaoli, L., Dudley, E.G., and Kovac, J.
689 (2019). *Bacillus clarus* sp. nov. is a new *Bacillus cereus* group species isolated from soil.
690 *bioRxiv*, 508077.

691 Akamatsu, R., Suzuki, M., Okinaka, K., Sasahara, T., Yamane, K., Suzuki, S., Fujikura, D.,
692 Furuta, Y., Ohnishi, N., Esaki, M., Shibayama, K., and Higashi, H. (2019). Novel
693 Sequence Type in *Bacillus cereus* Strains Associated with Nosocomial Infections and
694 Bacteremia, Japan. *Emerg Infect Dis* 25, 883-890.

695 Angiuoli, S.V., and Salzberg, S.L. (2011). Mugsy: fast multiple alignment of closely related
696 whole genomes. *Bioinformatics* 27, 334-342.

697 Antonation, K.S., Grutzmacher, K., Dupke, S., Mabon, P., Zimmermann, F., Lankester, F.,
698 Peller, T., Feistner, A., Todd, A., Herbinger, I., De Nys, H.M., Muyembe-Tamfun, J.J.,
699 Karhemere, S., Wittig, R.M., Couacy-Hymann, E., Grunow, R., Calvignac-Spencer, S.,
700 Corbett, C.R., Klee, S.R., and Leendertz, F.H. (2016). *Bacillus cereus* Biovar Anthracis
701 Causing Anthrax in Sub-Saharan Africa-Chromosomal Monophyly and Broad
702 Geographic Distribution. *PLoS Negl Trop Dis* 10, e0004923.

703 Arevalo, P., Vaninsberghe, D., Elsherbini, J., Gore, J., and Polz, M.F. (2019). A Reverse
704 Ecology Approach Based on a Biological Definition of Microbial Populations. *Cell* 178,
705 820-834 e814.

706 Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P.,
707 Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L.,
708 Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and

Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25**,** 25-29.

Avashia, S.B., Riggins, W.S., Lindley, C., Hoffmaster, A., Drumgoole, R., Nekomoto, T., Jackson, P.J., Hill, K.K., Williams, K., Lehman, L., Libal, M.C., Wilkins, P.P., Alexander, J., Tvaryanas, A., and Betz, T. (2007). Fatal pneumonia among metalworkers due to inhalation exposure to *Bacillus cereus* Containing *Bacillus anthracis* toxin genes. *Clin Infect Dis* 44**,** 414-416.

Brezillon, C., Haustant, M., Dupke, S., Corre, J.P., Lander, A., Franz, T., Monot, M., Couture-Tosi, E., Jouvion, G., Leendertz, F.H., Grunow, R., Mock, M.E., Klee, S.R., and Goossens, P.L. (2015). Capsules, toxins and AtxA as virulence factors of emerging *Bacillus cereus* biovar anthracis. *PLoS Negl Trop Dis* 9**,** e0003455.

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.

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.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10**,** 421.

Candela, T., and Fouet, A. (2006). Poly-gamma-glutamate in bacteria. *Mol Microbiol* 60**,** 1091-1098.

730 Cardazzo, B., Negrisolo, E., Carraro, L., Alberghini, L., Patarnello, T., and Giaccone, V. (2008).
731 Multiple-locus sequence typing and analysis of toxin genes in *Bacillus cereus* food-borne
732 isolates. *Appl Environ Microbiol* 74**,** 850-860.
733 Carroll, L.M., Kovac, J., Miller, R.A., and Wiedmann, M. (2017). Rapid, High-Throughput
734 Identification of Anthrax-Causing and Emetic *Bacillus cereus* Group Genome
735 Assemblies via BTyper, a Computational Tool for Virulence-Based Classification of
736 *Bacillus cereus* Group Isolates by Using Nucleotide Sequencing Data. *Appl Environ*
737 *Microbiol* 83.
738 Carroll, L.M., and Wiedmann, M. (2020). Cereulide synthetase acquisition and loss events
739 within the evolutionary history of Group III *Bacillus cereus sensu lato* facilitate the
740 transition between emetic and diarrheal foodborne pathogen. *bioRxiv***,**
741 2020.2005.2012.090951.
742 Carroll, L.M., Wiedmann, M., and Kovac, J. (2020). Proposal of a Taxonomic Nomenclature for
743 the *Bacillus cereus* Group Which Reconciles Genomic Definitions of Bacterial Species
744 with Clinical and Industrial Phenotypes. *mBio* 11.
745 Carroll, L.M., Wiedmann, M., Mukherjee, M., Nicholas, D.C., Mingle, L.A., Dumas, N.B., Cole,
746 J.A., and Kovac, J. (2019). Characterization of Emetic and Diarrheal *Bacillus cereus*
747 Strains From a 2016 Foodborne Outbreak Using Whole-Genome Sequencing: Addressing
748 the Microbiological, Epidemiological, and Bioinformatic Challenges. *Front Microbiol*
749 10**,** 144.
750 Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research.
751 *InterJournal, Complex Systems 1695*.

752 Durak, M.Z., Fromm, H.I., Huck, J.R., Zadoks, R.N., and Boor, K.J. (2006). Development of
753 Molecular Typing Methods for *Bacillus* spp. and *Paenibacillus* spp. Isolated from Fluid
754 Milk Products. *Journal of Food Science* 71**,** M50-M56.
755 Ehling-Schulz, M., Frenzel, E., and Gohar, M. (2015). Food-bacteria interplay: pathometabolism
756 of emetic *Bacillus cereus*. *Front Microbiol* 6**,** 704.
757 Ehling-Schulz, M., Fricker, M., Grallert, H., Rieck, P., Wagner, M., and Scherer, S. (2006).
758 Cereulide synthetase gene cluster from emetic *Bacillus cereus*: structure and location on a
759 mega virulence plasmid related to *Bacillus anthracis* toxin plasmid pXO1. *BMC*
760 *Microbiol* 6**,** 20.
761 Ehling-Schulz, M., Lereclus, D., and Koehler, T.M. (2019). The *Bacillus cereus* Group: *Bacillus*
762 Species with Pathogenic Potential. *Microbiol Spectr* 7.
763 Ehling-Schulz, M., Svensson, B., Guinebretiere, M.H., Lindback, T., Andersson, M., Schulz, A.,
764 Fricker, M., Christiansson, A., Granum, P.E., Martlbauer, E., Nguyen-The, C., Salkinoja-
765 Salonen, M., and Scherer, S. (2005). Emetic toxin formation of *Bacillus cereus* is
766 restricted to a single evolutionary lineage of closely related strains. *Microbiology* 151**,**
767 183-197.
768 Elshaghabee, F.M.F., Rokana, N., Gulhane, R.D., Sharma, C., and Panwar, H. (2017). *Bacillus*
769 As Potential Probiotics: Status, Concerns, and Future Perspectives. *Front Microbiol* 8**,**
770 1490.
771 Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome
772 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 16**,** 157.
773 Fagerlund, A., Lindback, T., and Granum, P.E. (2010). *Bacillus cereus* cytotoxins Hbl, Nhe and
774 CytK are secreted via the Sec translocation pathway. *BMC Microbiol* 10**,** 304.

775 Fagerlund, A., Ween, O., Lund, T., Hardy, S.P., and Granum, P.E. (2004). Genetic and
776 functional analysis of the *cytK* family of genes in *Bacillus cereus*. *Microbiology* 150**,**
777 2689-2697.
778 Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-
779 generation sequencing data. *Bioinformatics* 28**,** 3150-3152.
780 Galili, T. (2015). dendextend: an R package for visualizing, adjusting and comparing trees of
781 hierarchical clustering. *Bioinformatics* 31**,** 3718-3720.
782 Gardner, S.N., and Hall, B.G. (2013). When whole-genome alignments just won't work: kSNP
783 v2 software for alignment-free SNP discovery and phylogenetics of hundreds of
784 microbial genomes. *PLoS One* 8**,** e81760.
785 Gardner, S.N., Slezak, T., and Hall, B.G. (2015). kSNP3.0: SNP detection and phylogenetic
786 analysis of genomes without genome alignment or reference genome. *Bioinformatics* 31**,**
787 2877-2878.
788 Garnier, S. (2018). "viridis: Default Color Maps from 'matplotlib'". 0.5.1 ed..
789 Gdoura-Ben Amor, M., Siala, M., Zayani, M., Grosset, N., Smaoui, S., Messadi-Akrout, F.,
790 Baron, F., Jan, S., Gautier, M., and Gdoura, R. (2018). Isolation, Identification,
791 Prevalence, and Genetic Diversity of *Bacillus cereus* Group Bacteria From Different
792 Foodstuffs in Tunisia. *Front Microbiol* 9**,** 447.
793 Glasset, B., Herbin, S., Granier, S.A., Cavalie, L., Lafeuille, E., Guerin, C., Ruimy, R.,
794 Casagrande-Magne, F., Levast, M., Chautemps, N., Decousser, J.W., Belotti, L., Pelloux,
795 I., Robert, J., Brisabois, A., and Ramarao, N. (2018). *Bacillus cereus*, a serious cause of
796 nosocomial infections: Epidemiologic and genetic survey. *PLoS One* 13**,** e0194346.

797 Glasset, B., Herbin, S., Guillier, L., Cadel-Six, S., Vignaud, M.L., Grout, J., Pairaud, S., Michel,
798 V., Hennekinne, J.A., Ramarao, N., and Brisabois, A. (2016). *Bacillus cereus*-induced
799 food-borne outbreaks in France, 2007 to 2014: epidemiology and genetic
800 characterisation. *Euro Surveill* 21.
801 Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., and Tiedje, J.M.
802 (2007). DNA-DNA hybridization values and their relationship to whole-genome
803 sequence similarities. *Int J Syst Evol Microbiol* 57**,** 81-91.
804 Guinebretiere, M.-H., Fagerlund, A., Granum, P.E., and Nguyen-The, C. (2006). Rapid
805 discrimination of *cytK-1* and *cytK-2* genes in *Bacillus cereus* strains by a novel duplex
806 PCR system. *FEMS Microbiology Letters* 259**,** 74-80.
807 Guinebretiere, M.H., Auger, S., Galleron, N., Contzen, M., De Sarrau, B., De Buyser, M.L.,
808 Lamberet, G., Fagerlund, A., Granum, P.E., Lereclus, D., De Vos, P., Nguyen-The, C.,
809 and Sorokin, A. (2013). *Bacillus cytotoxicus* sp. nov. is a novel thermotolerant species of
810 the *Bacillus cereus* Group occasionally associated with food poisoning. *Int J Syst Evol*
811 *Microbiol* 63**,** 31-40.
812 Guinebretiere, M.H., Thompson, F.L., Sorokin, A., Normand, P., Dawyndt, P., Ehling-Schulz,
813 M., Svensson, B., Sanchis, V., Nguyen-The, C., Heyndrickx, M., and De Vos, P. (2008).
814 Ecological diversification in the *Bacillus cereus* Group. *Environ Microbiol* 10**,** 851-865.
815 Guinebretiere, M.H., Velge, P., Couvert, O., Carlin, F., Debuyser, M.L., and Nguyen-The, C.
816 (2010). Ability of *Bacillus cereus* group strains to cause food poisoning varies according
817 to phylogenetic affiliation (groups I to VII) rather than species affiliation. *J Clin*
818 *Microbiol* 48**,** 3388-3391.

819 Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool
820 for genome assemblies. *Bioinformatics* 29**,** 1072-1075.
821 Hanage, W.P., Fraser, C., and Spratt, B.G. (2005). Fuzzy species among recombinogenic
822 bacteria. *BMC Biol* 3**,** 6.
823 Hoang, D.T., Chernomor, O., Von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2:
824 Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* 35**,** 518-522.
825 Hoffmaster, A.R., Ravel, J., Rasko, D.A., Chapman, G.D., Chute, M.D., Marston, C.K., De,
826 B.K., Sacchi, C.T., Fitzgerald, C., Mayer, L.W., Maiden, M.C., Priest, F.G., Barker, M.,
827 Jiang, L., Cer, R.Z., Rilstone, J., Peterson, S.N., Weyant, R.S., Galloway, D.R., Read,
828 T.D., Popovic, T., and Fraser, C.M. (2004). Identification of anthrax toxin genes in a
829 *Bacillus cereus* associated with an illness resembling inhalation anthrax. *Proc Natl Acad*
830 *Sci U S A* 101**,** 8449-8454.
831 Huck, J.R., Hammond, B.H., Murphy, S.C., Woodcock, N.H., and Boor, K.J. (2007a). Tracking
832 spore-forming bacterial contaminants in fluid milk-processing systems. *J Dairy Sci* 90**,**
833 4872-4883.
834 Huck, J.R., Woodcock, N.H., Ralyea, R.D., and Boor, K.J. (2007b). Molecular subtyping and
835 characterization of psychrotolerant endospore-forming bacteria in two New York State
836 fluid milk processing systems. *J Food Prot* 70**,** 2354-2364.
837 Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., Von Mering, C., and
838 Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology
839 Assignment by eggNOG-Mapper. *Mol Biol Evol* 34**,** 2115-2122.
840 Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernandez-Plaza, A., Forslund, S.K., Cook, H.,
841 Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., Von Mering, C., and Bork, P. (2019).

842 eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology
843 resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 47**,** D309-D314.
844 Huxley, J. (1943). Systematics and the origin of Species: from the Viewpoint of a Zoologist.
845 *Nature* 151**,** 347-348.
846 Ivy, R.A., Ranieri, M.L., Martin, N.H., Den Bakker, H.C., Xavier, B.M., Wiedmann, M., and
847 Boor, K.J. (2012). Identification and characterization of psychrotolerant sporeformers
848 associated with fluid milk production and processing. *Appl Environ Microbiol* 78**,** 1853-
849 1864.
850 Jain, C., Rodriguez, R.L., Phillippy, A.M., Konstantinidis, K.T., and Aluru, S. (2018). High
851 throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries.
852 *Nat Commun* 9**,** 5114.
853 Jessberger, N., Krey, V.M., Rademacher, C., Bohm, M.E., Mohr, A.K., Ehling-Schulz, M.,
854 Scherer, S., and Martlbauer, E. (2015). From genome to toxicity: a combinatory approach
855 highlights the complexity of enterotoxin production in *Bacillus cereus*. *Front Microbiol*
856 6**,** 560.
857 Jimenez, G., Urdiain, M., Cifuentes, A., Lopez-Lopez, A., Blanch, A.R., Tamames, J., Kampfer,
858 P., Kolsto, A.B., Ramon, D., Martinez, J.F., Codoner, F.M., and Rossello-Mora, R.
859 (2013). Description of *Bacillus toyonensis* sp. nov., a novel species of the *Bacillus cereus*
860 group, and pairwise genome comparisons of the species of the group by means of ANI
861 calculations. *Syst Appl Microbiol* 36**,** 383-391.
862 Johler, S., Kalbhenn, E.M., Heini, N., Brodmann, P., Gautsch, S., Bagcioglu, M., Contzen, M.,
863 Stephan, R., and Ehling-Schulz, M. (2018). Enterotoxin Production of *Bacillus*

thuringiensis Isolates From Biopesticides, Foods, and Outbreaks. *Front Microbiol* 9**,** 1915.

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.

Jolley, K.A., and Maiden, M.C. (2010). BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 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.

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.

Jung, M.Y., Kim, J.S., Paek, W.K., Lim, J., Lee, H., Kim, P.I., Ma, J.Y., Kim, W., and Chang, Y.H. (2011). *Bacillus manliponensis* sp. nov., a new member of the *Bacillus cereus* group isolated from foreshore tidal flat sediment. *J Microbiol* 49**,** 1027-1032.

Jung, M.Y., Paek, W.K., Park, I.S., Han, J.R., Sin, Y., Paek, J., Rhee, M.S., Kim, H., Song, H.S., and Chang, Y.H. (2010). *Bacillus gaemokensis* sp. nov., isolated from foreshore tidal flat sediment from the Yellow Sea. *J Microbiol* 48**,** 867-871.

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.

885 Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid
886 multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30**,**
887 3059-3066.
888 Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7:
889 improvements in performance and usability. *Mol Biol Evol* 30**,** 772-780.
890 Katz, L.S., Griswold, T., Williams-Newkirk, A.J., Wagner, D., Petkau, A., Sieffert, C., Van
891 Domselaar, G., Deng, X., and Carleton, H.A. (2017). A Comparative Analysis of the
892 Lyve-SET Phylogenomics Pipeline for Genomic Epidemiology of Foodborne Pathogens.
893 *Front Microbiol* 8**,** 375.
894 Kendall, M., and Colijn, C. (2015). A tree metric using structure and length to capture distinct
895 phylogenetic signals. *arXiv***,** 1507.05211.
896 Kendall, M., and Colijn, C. (2016). Mapping Phylogenetic Trees to Reveal Distinct Patterns of
897 Evolution. *Molecular Biology and Evolution* 33**,** 2735-2743.
898 Kindle, P., Etter, D., Stephan, R., and Johler, S. (2019). Population structure and toxin gene
899 profiles of *Bacillus cereus sensu lato* isolated from flour products. *FEMS Microbiol Lett*
900 366.
901 Klee, S.R., Brzuszkiewicz, E.B., Nattermann, H., Bruggemann, H., Dupke, S., Wollherr, A.,
902 Franz, T., Pauli, G., Appel, B., Liebl, W., Couacy-Hymann, E., Boesch, C., Meyer, F.D.,
903 Leendertz, F.H., Ellerbrok, H., Gottschalk, G., Grunow, R., and Liesegang, H. (2010).
904 The genome of a *Bacillus* isolate causing anthrax in chimpanzees combines chromosomal
905 properties of *B. cereus* with *B. anthracis* virulence plasmids. *PLoS One* 5**,** e10986.

906 Kovac, J., Miller, R.A., Carroll, L.M., Kent, D.J., Jian, J., Beno, S.M., and Wiedmann, M.
907 (2016). Production of hemolysin BL by *Bacillus cereus* group isolates of dairy origin is
908 associated with whole-genome phylogenetic clade. *BMC Genomics* 17**,** 581.
909 Kruskal, J.B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*
910 29**,** 115-129.
911 Lazarte, J.N., Lopez, R.P., Ghiringhelli, P.D., and Beron, C.M. (2018). *Bacillus wiedmannii*
912 biovar thuringiensis: A Specialized Mosquitocidal Pathogen with Plasmids from Diverse
913 Origins. *Genome Biol Evol* 10**,** 2823-2833.
914 Lechner, S., Mayr, R., Francis, K.P., Pruss, B.M., Kaplan, T., Wiessner-Gunkel, E., Stewart,
915 G.S., and Scherer, S. (1998). *Bacillus weihenstephanensis* sp. nov. is a new
916 psychrotolerant species of the *Bacillus cereus* group. *Int J Syst Bacteriol* 48 Pt 4**,** 1373-
917 1382.
918 Lee, I., Ouk Kim, Y., Park, S.C., and Chun, J. (2016). OrthoANI: An improved algorithm and
919 software for calculating average nucleotide identity. *Int J Syst Evol Microbiol* 66**,** 1100-
920 1103.
921 Leendertz, F.H., Ellerbrok, H., Boesch, C., Couacy-Hymann, E., Matz-Rensing, K., Hakenbeck,
922 R., Bergmann, C., Abaza, P., Junglen, S., Moebius, Y., Vigilant, L., Formenty, P., and
923 Pauli, G. (2004). Anthrax kills wild chimpanzees in a tropical rainforest. *Nature* 430**,**
924 451-452.
925 Lewis, P.O. (2001). A likelihood approach to estimating phylogeny from discrete morphological
926 character data. *Syst Biol* 50**,** 913-925.
927 Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of
928 protein or nucleotide sequences. *Bioinformatics* 22**,** 1658-1659.

929 Liu, B., Liu, G.H., Hu, G.P., Sengonca, C., Lin, N.Q., Tang, J.Y., Tang, W.Q., and Lin, Y.Z.
930 (2014). *Bacillus bingmayongensis* sp. nov., isolated from the pit soil of Emperor Qin's
931 Terra-cotta warriors in China. *Antonie Van Leeuwenhoek* 105**,** 501-510.
932 Liu, Y., Du, J., Lai, Q., Zeng, R., Ye, D., Xu, J., and Shao, Z. (2017). Proposal of nine novel
933 species of the *Bacillus cereus* group. *Int J Syst Evol Microbiol* 67**,** 2499-2508.
934 Maechler, M., Rousseeuw, P., Struyf, A., Hubert , M., and Hornik, K. (2019). "cluster: Cluster
935 Analysis Basics and Extensions". 2.1.0 ed.).
936 Marston, C.K., Ibrahim, H., Lee, P., Churchwell, G., Gumke, M., Stanek, D., Gee, J.E., Boyer,
937 A.E., Gallegos-Candela, M., Barr, J.R., Li, H., Boulay, D., Cronin, L., Quinn, C.P., and
938 Hoffmaster, A.R. (2016). Anthrax Toxin-Expressing *Bacillus cereus* Isolated from an
939 Anthrax-Like Eschar. *PLoS One* 11**,** e0156987.
940 Messelhäußer, U., and Ehling-Schulz, M. (2018). *Bacillus cereus*—a Multifaceted Opportunistic
941 Pathogen. *Current Clinical Microbiology Reports* 5**,** 120-125.
942 Miller, R.A., Beno, S.M., Kent, D.J., Carroll, L.M., Martin, N.H., Boor, K.J., and Kovac, J.
943 (2016). *Bacillus wiedmannii* sp. nov., a psychrotolerant and cytotoxic *Bacillus cereus*
944 group species isolated from dairy foods and dairy environments. *Int J Syst Evol Microbiol*
945 66**,** 4744-4753.
946 Miller, R.A., Jian, J., Beno, S.M., Wiedmann, M., and Kovac, J. (2018). Intraclade Variability in
947 Toxin Production and Cytotoxicity of *Bacillus cereus* Group Type Strains and Dairy-
948 Associated Isolates. *Appl Environ Microbiol* 84.
949 Minh, B.Q., Nguyen, M.A., and Von Haeseler, A. (2013). Ultrafast approximation for
950 phylogenetic bootstrap. *Mol Biol Evol* 30**,** 1188-1195.

951 Moayeri, M., Leppla, S.H., Vrentas, C., Pomerantsev, A.P., and Liu, S. (2015). Anthrax
952 Pathogenesis. *Annu Rev Microbiol* 69**,** 185-208.
953 Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and
954 effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol*
955 *Evol* 32**,** 268-274.
956 Oh, S.Y., Budzik, J.M., Garufi, G., and Schneewind, O. (2011). Two capsular polysaccharides
957 enable *Bacillus cereus* G9241 to cause anthrax-like disease. *Mol Microbiol* 80**,** 455-470.
958 Okinaka, R.T., Cloud, K., Hampton, O., Hoffmaster, A.R., Hill, K.K., Keim, P., Koehler, T.M.,
959 Lamke, G., Kumano, S., Mahillon, J., Manter, D., Martinez, Y., Ricke, D., Svensson, R.,
960 and Jackson, P.J. (1999). Sequence and organization of pXO1, the large *Bacillus*
961 *anthracis* plasmid harboring the anthrax toxin genes. *J Bacteriol* 181**,** 6509-6515.
962 Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., Mcglinn, D., Minchin, P.R.,
963 O'hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., and Wagner, H.
964 (2019). "vegan: Community Ecology Package. R package version 2.5-6. https://CRAN.R-
965 project.org/package=vegan".).
966 Ozdemir, F., and Arslan, S. (2019). Molecular Characterization and Toxin Profiles of *Bacillus*
967 spp. Isolated from Retail Fish and Ground Beef. *J Food Sci* 84**,** 548-556.
968 Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution
969 in R language. *Bioinformatics* 20**,** 289-290.
970 Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and
971 evolutionary analyses in R. *Bioinformatics* 35**,** 526-528.

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.

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

Pilo, P., and Frey, J. (2018). Pathogenicity, population genetics and dissemination of *Bacillus anthracis*. *Infect Genet Evol* 64**,** 115-125.

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. *Analytical Methods* 8**,** 12-24.

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**,** D61-65.

R Core Team (2019). "R: A Language and Environment for Statistical Computing". 3.6.1 ed. (Vienna, Austria: R Foundation for Statistical Computing).

R Hackathon (2019). "phylobase: Base Package for Phylogenetic Structures and Comparative Data". 0.8.6 ed.).

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.

Revell, L.J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 3**,** 217-223.

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.

995 Richter, M., Rossello-Mora, R., Oliver Glockner, F., and Peplies, J. (2016). JSpeciesWS: a web
996 server for prokaryotic species circumscription based on pairwise genome comparison.
997 *Bioinformatics* 32**,** 929-931.
998 Rigaux, C., Ancelet, S., Carlin, F., Nguyen-The, C., and Albert, I. (2013). Inferring an
999 augmented Bayesian network to confront a complex quantitative microbial risk
1000 assessment model with durability studies: application to *Bacillus cereus* on a courgette
1001 puree production chain. *Risk Anal* 33**,** 877-892.
1002 Riol, C.D., Dietrich, R., Martlbauer, E., and Jessberger, N. (2018). Consumed Foodstuffs Have a
1003 Crucial Impact on the Toxic Activity of Enteropathogenic *Bacillus cereus*. *Front*
1004 *Microbiol* 9**,** 1946.
1005 Rocha, E.P.C. (2018). Neutral Theory, Microbial Practice: Challenges in Bacterial Population
1006 Genetics. *Mol Biol Evol* 35**,** 1338-1347.
1007 Rodriguez, R.L., Gunturu, S., Harvey, W.T., Rossello-Mora, R., Tiedje, J.M., Cole, J.R., and
1008 Konstantinidis, K.T. (2018). The Microbial Genomes Atlas (MiGA) webserver:
1009 taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome
1010 level. *Nucleic Acids Res* 46**,** W282-W288.
1011 Romero-Alvarez, D., Peterson, A.T., Salzer, J.S., Pittiglio, C., Shadomy, S., Traxler, R., Vieira,
1012 A.R., Bower, W.A., Walke, H., and Campbell, L.P. (2020). Potential distributions of
1013 *Bacillus anthracis* and *Bacillus cereus* biovar anthracis causing anthrax in Africa. *PLoS*
1014 *Negl Trop Dis* 14**,** e0008131.
1015 Rosvall, M., Axelsson, D., and Bergstrom, C.T. (2009). The map equation. *The European*
1016 *Physical Journal Special Topics* 178**,** 13-23.

1017 Rouzeau-Szynalski, K., Stollewerk, K., Messelhausser, U., and Ehling-Schulz, M. (2020). Why
1018 be serious about emetic *Bacillus cereus*: Cereulide production and industrial challenges.
1019 *Food Microbiol* 85**,** 103279.
1020 Schoeni, J.L., and Wong, A.C. (2005). *Bacillus cereus* food poisoning and its toxins. *J Food Prot*
1021 68**,** 636-648.
1022 Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30**,** 2068-
1023 2069.
1024 Senesi, S., and Ghelardi, E. (2010). Production, secretion and biological activity of *Bacillus*
1025 *cereus* enterotoxins. *Toxins (Basel)* 2**,** 1690-1703.
1026 Stenfors Arnesen, L.P., Fagerlund, A., and Granum, P.E. (2008). From soil to gut: *Bacillus*
1027 *cereus* and its food poisoning toxins. *FEMS Microbiol Rev* 32**,** 579-606.
1028 Stevens, M.J.A., Tasara, T., Klumpp, J., Stephan, R., Ehling-Schulz, M., and Johler, S. (2019).
1029 Whole-genome-based phylogeny of *Bacillus cytotoxicus* reveals different clades within
1030 the species and provides clues on ecology and evolution. *Sci Rep* 9**,** 1984.
1031 Tallent, S.M., Knolhoff, A., Rhodehamel, E.J., Harmon, S.M., and Bennett, R.W. (2019).
1032 "Chapter 14: *Bacillus cereus*," in *Bacteriological Analytical Manual (BAM)*. 8th ed: Food
1033 and Drug Administration).
1034 Tallent, S.M., Kotewicz, K.M., Strain, E.A., and Bennett, R.W. (2012). Efficient isolation and
1035 identification of *Bacillus cereus* group. *J AOAC Int* 95**,** 446-451.
1036 Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences.
1037 *Lectures on mathematics in the life sciences* 17**,** 57-86.
1038 The gene ontology consortium (2018). The Gene Ontology Resource: 20 years and still GOing
1039 strong. *Nucleic Acids Research* 47**,** D330-D338.

1040 Thorsen, L., Kando, C.K., Sawadogo, H., Larsen, N., Diawara, B., Ouedraogo, G.A.,
1041 Hendriksen, N.B., and Jespersen, L. (2015). Characteristics and phylogeny of *Bacillus*
1042 *cereus* strains isolated from Maari, a traditional West African food condiment. *Int J Food*
1043 *Microbiol* 196**,** 70-78.
1044 Tonkin-Hill, G., Lees, J.A., Bentley, S.D., Frost, S.D.W., and Corander, J. (2018). RhierBAPS:
1045 An R implementation of the population clustering algorithm hierBAPS. *Wellcome Open*
1046 *Res* 3**,** 93.
1047 Webb, M.D., Barker, G.C., Goodburn, K.E., and Peck, M.W. (2019). Risk presented to
1048 minimally processed chilled foods by psychrotrophic *Bacillus cereus*. *Trends Food Sci*
1049 *Technol* 93**,** 94-105.
1050 Wickham, H. (2007). Reshaping Data with the reshape Package. *2007* 21**,** 20.
1051 Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.
1052 Wickham, H., and Bryan, J. (2019). "readxl: Read Excel Files". 1.3.1 ed.).
1053 Wickham, H., François, R., Henry, L., and Müller, K. (2020). "dplyr: A Grammar of Data
1054 Manipulation". 0.8.5 ed.).
1055 Wilson, M.K., Vergis, J.M., Alem, F., Palmer, J.R., Keane-Myers, A.M., Brahmbhatt, T.N.,
1056 Ventura, C.L., and O'brien, A.D. (2011). *Bacillus cereus* G9241 makes anthrax toxin and
1057 capsule like highly virulent *B. anthracis* Ames but behaves like attenuated toxigenic
1058 nonencapsulated *B. anthracis* Sterne in rabbits and mice. *Infect Immun* 79**,** 3012-3019.
1059 Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with
1060 variable rates over sites: Approximate methods. *Journal of Molecular Evolution* 39**,** 306-
1061 314.

1062 Yoon, S.H., Ha, S.M., Lim, J., Kwon, S., and Chun, J. (2017). A large-scale evaluation of
1063 algorithms to calculate average nucleotide identity. *Antonie Van Leeuwenhoek* 110**,** 1281-
1064 1286.
1065 Yu, G., Lam, T.T., Zhu, H., and Guan, Y. (2018). Two Methods for Mapping and Visualizing
1066 Associated Data on Phylogeny Using Ggtree. *Mol Biol Evol* 35**,** 3041-3043.
1067 Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.-Y. (2017). ggtree: an r package for
1068 visualization and annotation of phylogenetic trees with their covariates and other
1069 associated data. *Methods in Ecology and Evolution* 8**,** 28-36.
1070 Zhao, C., and Wang, Z. (2018). GOGO: An improved algorithm to measure the semantic
1071 similarity between gene ontology terms. *Sci Rep* 8**,** 15107.
1072 Zhao, S., Chen, J., Fei, P., Feng, H., Wang, Y., Ali, M.A., Li, S., Jing, H., and Yang, W. (2020).
1073 Prevalence, molecular characterization, and antibiotic susceptibility of *Bacillus cereus*
1074 isolated from dairy products in China. *J Dairy Sci* 103**,** 3994-4001.
1075 Zhuang, K., Li, H., Zhang, Z., Wu, S., Zhang, Y., Fox, E.M., Man, C., and Jiang, Y. (2019).
1076 Typing and evaluating heat resistance of *Bacillus cereus sensu stricto* isolated from the
1077 processing environment of powdered infant formula. *J Dairy Sci* 102**,** 7781-7793.
1078
1079
1080

TABLES

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 ≳ 92.5 ANI with *B. pseudomycoides* str. DSM 12442^T (GCF_000161455.1)
B. mosaicus	Groups II/III	Groups II/III	Shares ≳ 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 ≳ 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 ≳ 92.5 ANI with *B. toyonensis* str. BCT-7112^T (GCF_000496285.1)
B. mycoides	Groups II/III/VI	Groups VI/VIII	Shares ≳ 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 ≳ 92.5 ANI with *B. cytotoxicus* str. NVH 391-98^T (GCF_000017425.1)
B. paramycoides	Group VI	Group VI	Shares ≳ 92.5 ANI with *B. paramycoides* str. NH24A2^T (GCF_001884235.1)
B. luti	Groups III/V/VI	Group II	Shares ≳ 92.5 ANI with *B. luti* str. TD41^T (GCF_001884105.1)

[a]See Supplementary Tables S5 and 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. (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. (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 Supplemental Table S7 of Carroll, et al. (Carroll et al., 2020) for a list of medoid genomes for the above genomospecies.

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 and 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

FIGURE LEGENDS

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/).

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.

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.

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 a phylogeny, using the optimal nucleotide substitution model selected using ModelFinder (i.e., the TVM+F+R4 model).

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,C, and 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

1172 locus. For each alignment, IQ-TREE v. 1.5.4 (A) and 1.6.5 (B, C, and D) was used to construct a
1173 phylogeny, using either the GTR+G+ASC nucleotide substitution model (A), or the optimal
1174 model selected using ModelFinder (B, C, and D).

1175 **Figure 6.** (A) Network of Clusters of Orthologous Groups (COG) functional categories assigned
1176 to 255 single-copy core genes that topologically mirror the *B. cereus s.l.* whole-genome
1177 phylogeny (Kendall-Colijn $P < 0.05$ after a Bonferroni correction). Each node corresponds to a
1178 COG functional category/group of functional categories assigned to one or more genes. Node
1179 size corresponds to the number of genes (out of 255 possible genes) assigned to a functional
1180 category/group of functional categories, ranging from one to 58 (for S, function unknown).
1181 Edges connect nodes that share one or more functional categories. Nodes of functional categories
1182 assigned to 15 or more genes are annotated with a text label denoting the number of genes
1183 assigned to the respective functional category. (B) Results of nonmetric multidimensional scaling
1184 (NMDS) performed using pairwise semantic/functional dissimilarities calculated between 94
1185 single-copy core genes based on their assigned Gene Ontology (GO) Biological Process
1186 Ontology (BPO) terms. Points represent individual genes, while shaded regions and convex hulls
1187 correspond to clusters of genes identified by GOGO, based on their BPO similarities. For a
1188 complete list of annotations associated with each of the 255 single-copy core genes, see
1189 Supplementary Table S2. For NMDS plots constructed using Cellular Component Ontology
1190 (CCO) and Molecular Function Ontology (MFO) dissimilarities, see Supplementary Figure S24.

1191 **Figure 7.** Distribution of selected *B. cereus s.l.* virulence factors within the *B. cereus s.l.*
1192 phylogeny (n = 1,741). Tip labels and branches within the phylogeny are colored by (A) *B.*
1193 *cereus s.l.* genomospecies, assigned using medoid genomes obtained at a 92.5 ANI threshold,
1194 and (B through 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. 2020 for detailed methods) (Carroll et al., 2020). 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.

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 and 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.

A. Construction of BTyper3 *B. cereus s.l.* Pseudo-Gene Flow Units

Step 0: Begin with 313 high-quality genomes that span *B. cereus s.l.*

= *B. cereus s.l.* genome

Step 1: Delineate true gene flow units among 313 genomes using PopCOGenT

= PopCOGent "Main Cluster"/true gene flow unit

Step 2: For each true gene flow unit identified by PopCOGenT, calculate pairwise ANI values between all genomes assigned to the gene flow unit

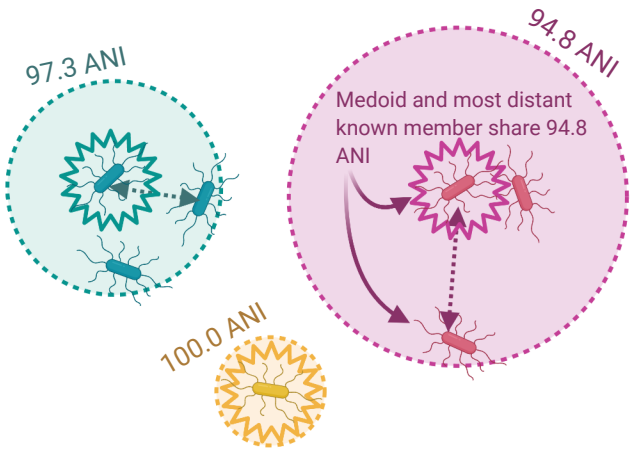
= Pairwise ANI value

Step 3: Identify the ANI-based medoid genome for each true PopCOGenT gene flow unit

= ANI-based medoid genome

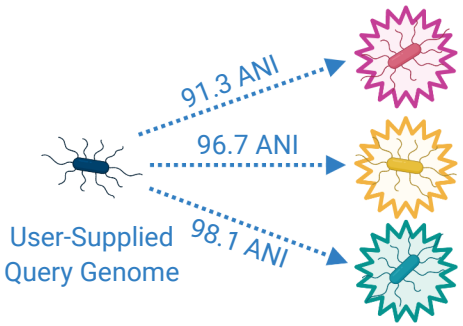
Step 4: Identify the genome in each true PopCOGenT gene flow unit that shares the lowest ANI value with the gene flow unit medoid genome; this is the observed ANI boundary for the resulting pseudo-gene flow unit

= Pseudo-gene flow unit

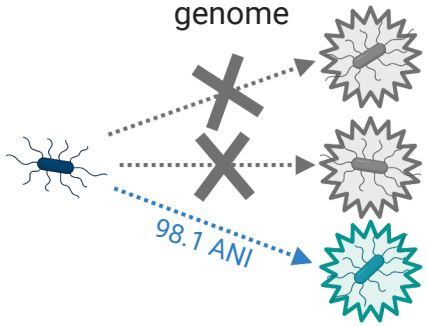


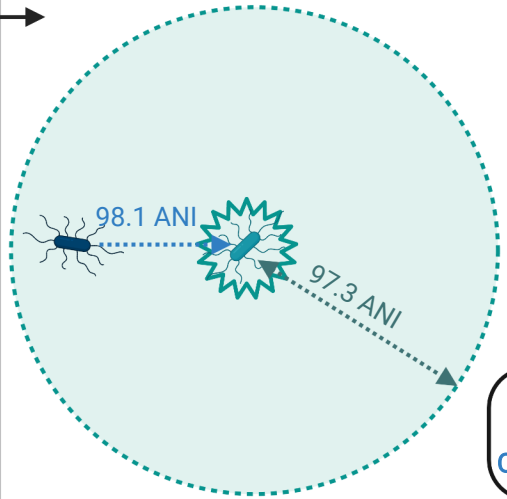
B. BTyper3 Pseudo-Gene Flow Unit Assignment Algorithm

Step 1: Calculate one-way ANI values between the query genome and all pseudo-gene flow unit medoid genomes



Step 2: Identify the pseudo-gene flow unit medoid genome that shares the highest ANI value with the query genome

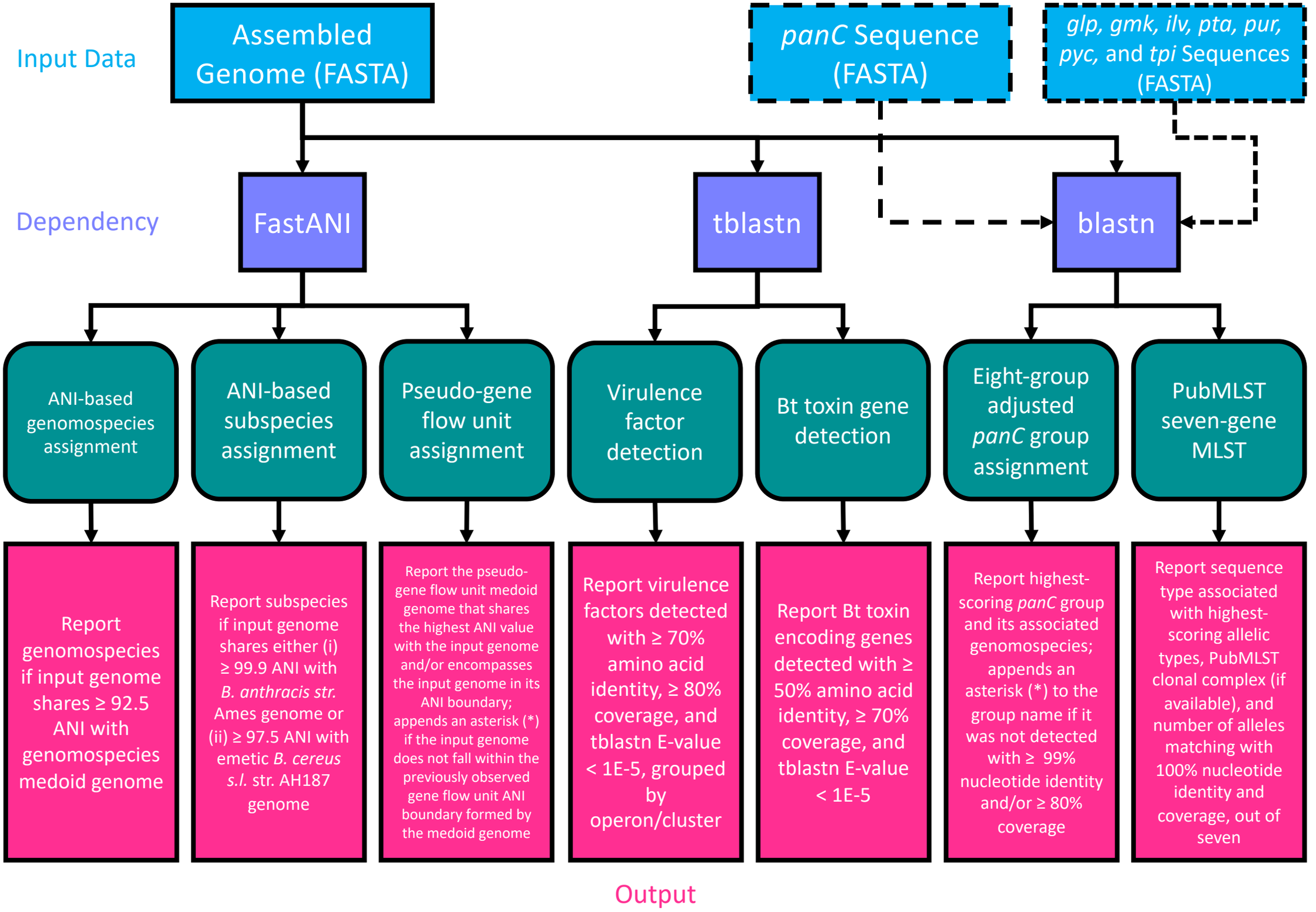


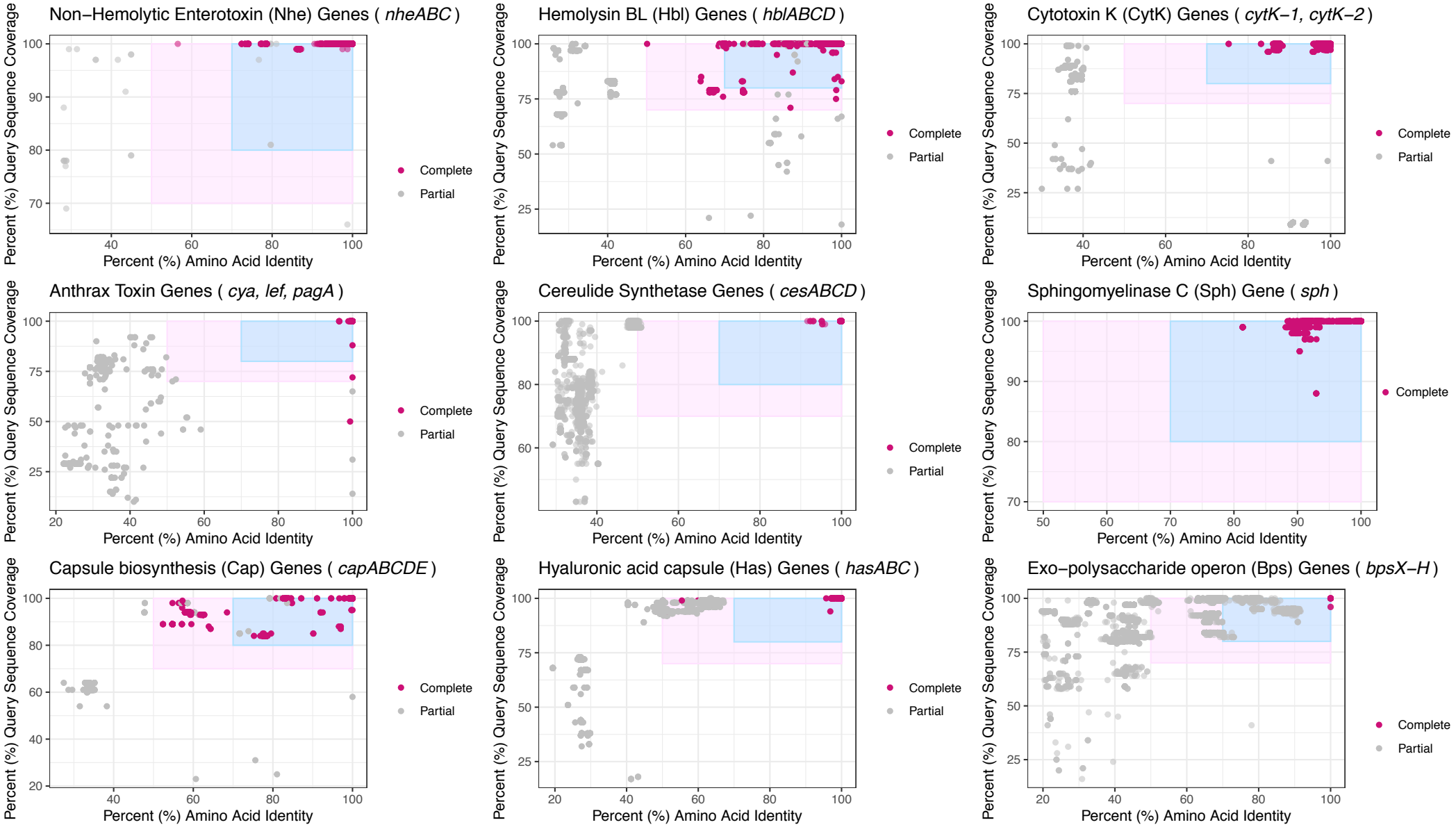


Step 3: If the ANI value shared between the query genome and the pseudo-gene flow unit medoid genome is within the observed ANI boundary for the pseudo-gene flow unit, the query genome is assinged to the pseudo-gene flow unit

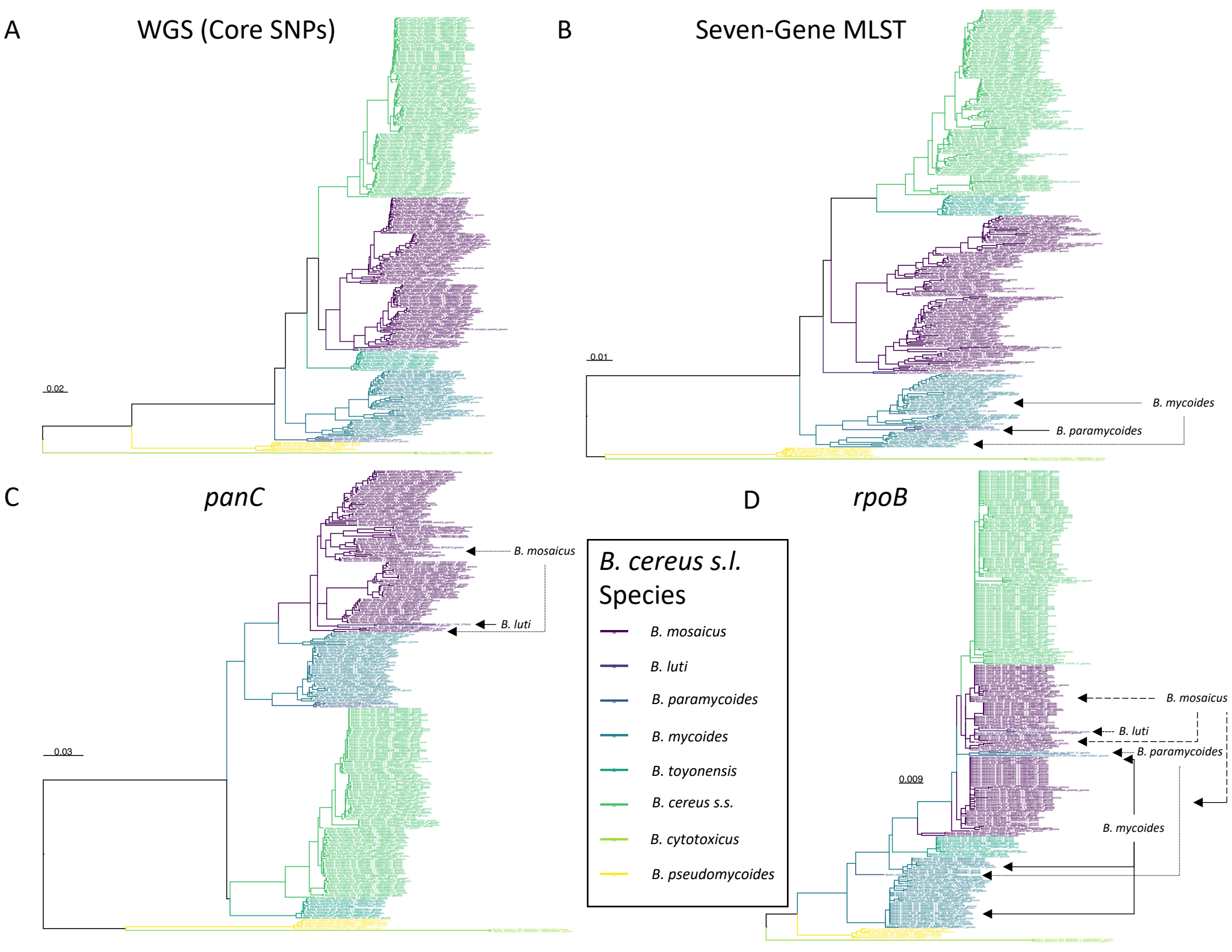
98.1 ≥ 97.3

Query genome assigned to **pseudo-gene flow unit**

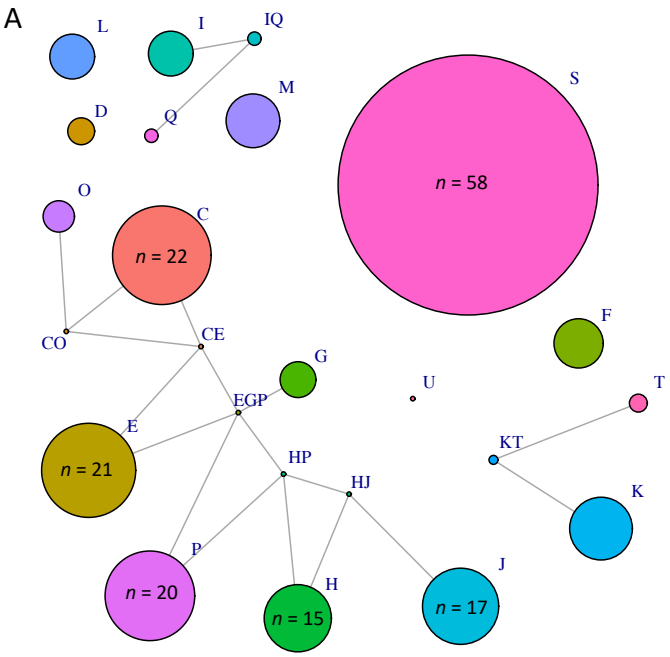

Input Data
Assembled Genome (FASTA)
panC Sequence (FASTA)
glp, gmk, ilv, pta, pur, pyc, and tpi Sequences (FASTA)
Dependency
FastANI
tblastn
blastn
ANI-based genomospecies assignment
ANI-based subspecies assignment
Pseudo-gene flow unit assignment
Virulence factor detection
Bt toxin gene detection
Eight-group adjusted panC group assignment
PubMLST seven-gene MLST
Report genomospecies if input genome shares ≥ 92.5 ANI with genomospecies medoid genome
Report subspecies if input genome shares either (i) ≥ 99.9 ANI with B. anthracis str. Ames genome or (ii) ≥ 97.5 ANI with emetic B. cereus s.l. str. AH187 genome
Report the pseudo-gene flow unit medoid genome that shares the highest ANI value with the input genome and/or encompasses the input genome in its ANI boundary; appends an asterisk (*) if the input genome does not fall within the previously observed gene flow unit ANI boundary formed by the medoid genome
Report virulence factors detected with ≥ 70% amino acid identity, ≥ 80% coverage, and tblastn E-value < 1E-5, grouped by operon/cluster
Report Bt toxin encoding genes detected with ≥ 50% amino acid identity, ≥ 70% coverage, and tblastn E-value < 1E-5
Report highest-scoring panC group and its associated genomospecies; appends an asterisk (*) to the group name if it was not detected with ≥ 99% nucleotide identity and/or ≥ 80% coverage
Report sequence type associated with highest-scoring allelic types, PubMLST clonal complex (if available), and number of alleles matching with 100% nucleotide identity and coverage, out of seven
Output


Non−Hemolytic Enterotoxin (Nhe) Genes (nheABC)
Hemolysin BL (Hbl) Genes (hblABCD)
Cytotoxin K (CytK) Genes (cytK−1, cytK−2)
Anthrax Toxin Genes (cya, lef, pagA)
Cereulide Synthetase Genes (cesABCD)
Sphingomyelinase C (Sph) Gene (sph)
Capsule biosynthesis (Cap) Genes (capABCDE)
Hyaluronic acid capsule (Has) Genes (hasABC)
Exo−polysaccharide operon (Bps) Genes (bpsX−H)
Percent (%) Query Sequence Coverage
Percent (%) Amino Acid Identity
Complete
Partial


A
Legacy panC Group I
Legacy panC Group V
Legacy panC Group II
Legacy panC Group VI
Legacy panC Group III
Legacy panC Group VII
Legacy panC Group IV
B
Adjusted panC Group I
Adjusted panC Group IVb
Adjusted panC Group VIII
Adjusted panC Group II
Adjusted panC Group V
Adjusted panC Group III
Adjusted panC Group VI
Adjusted panC Group IVa
Adjusted panC Group VII

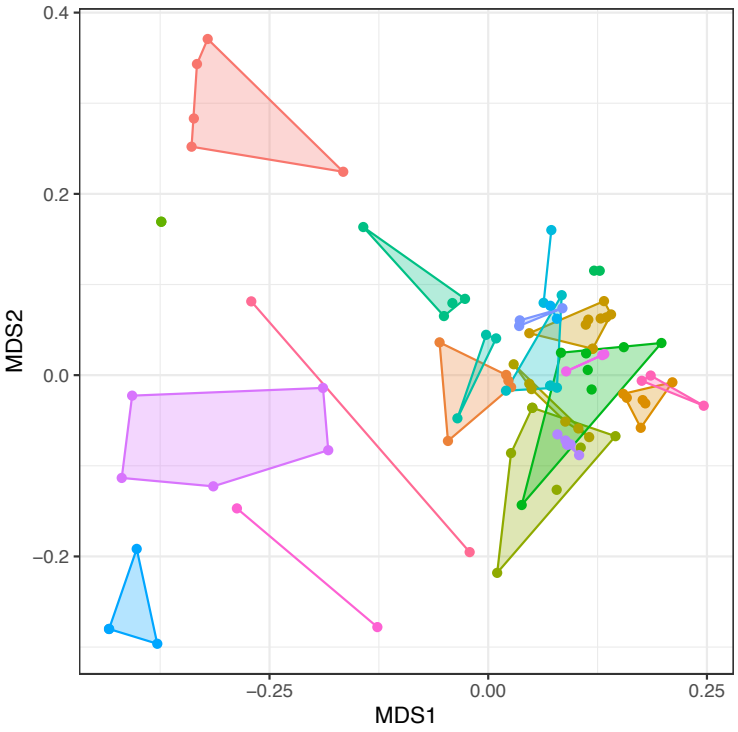

A
WGS (Core SNPs)
0.02
B
Seven-Gene MLST
0.01
B. mycoides
B. paramycoides
C
panC
B. mosaicus
B. luti
0.03
D
rpoB
B. mosaicus
B. luti
B. paramycoides
B. mycoides
0.009
B. cereus s.l.
Species
B. mosaicus
B. luti
B. paramycoides
B. mycoides
B. toyonensis
B. cereus s.s.
B. cytotoxicus
B. pseudomycoides

A



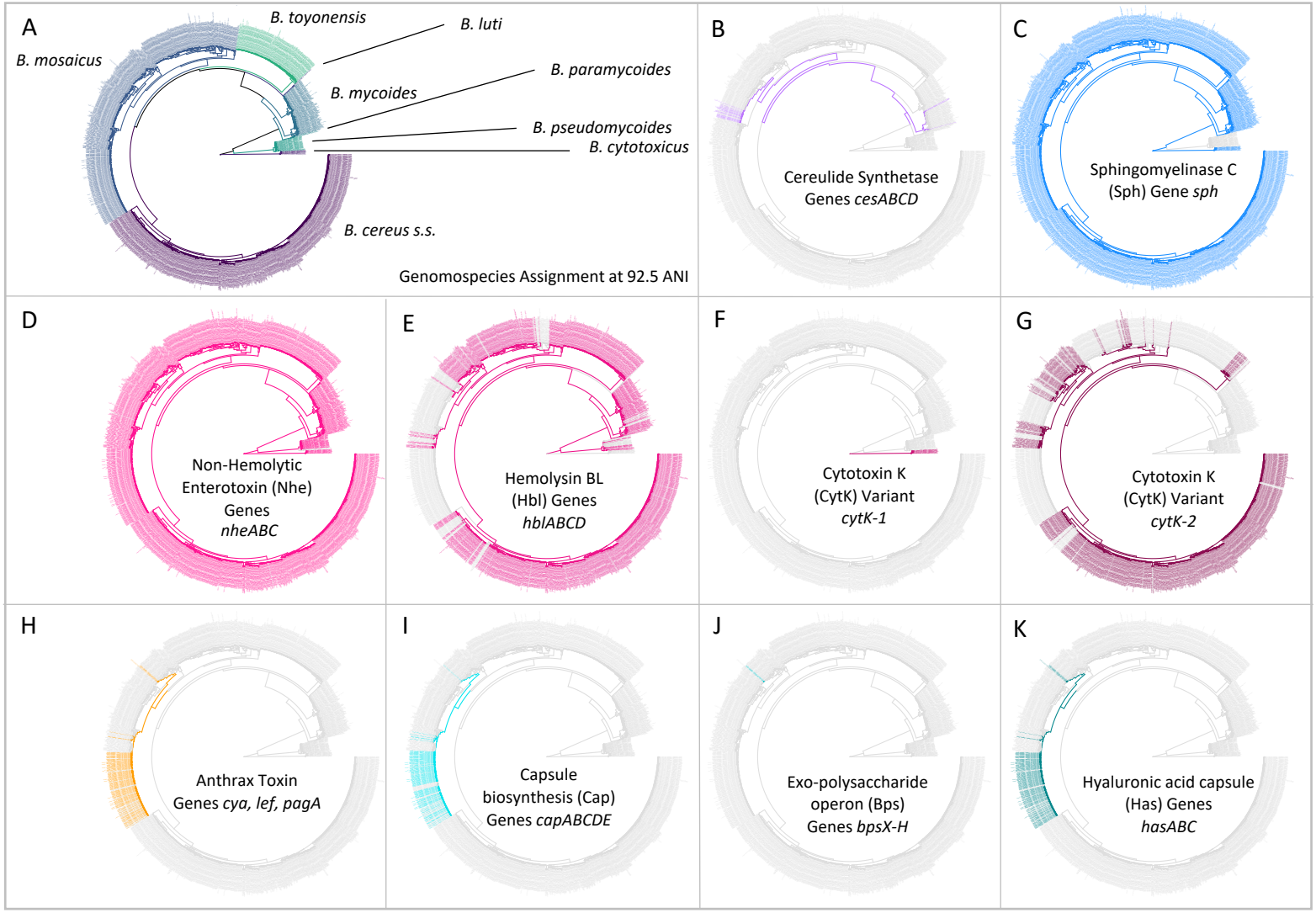
- C (energy production and conversion)
- D (cell cycle control, cell division, chromosome partitioning)
- E (amino acid metabolism and transport)
- F (nucleotide metabolism and transport)
- G (carbohydrate metabolism and transport)
- H (coenzyme metabolism and transport)
- I (lipid metabolism and transport)
- J (translation, ribosomal structure, and biogenesis)
- K (transcription)
- L (replication, recombination, repair)
- M (cell wall/membrane/envelope biogenesis)
- O (post-translational modification, protein turnover, chaperones)
- P (inorganic ion transport and metabolism)
- Q (secondary metabolite biosynthesis, transport, catabolism)
- S (function unknown)
- T (signal transduction mechanisms)
- U (intracellular trafficking, secretion, vesicular transport)

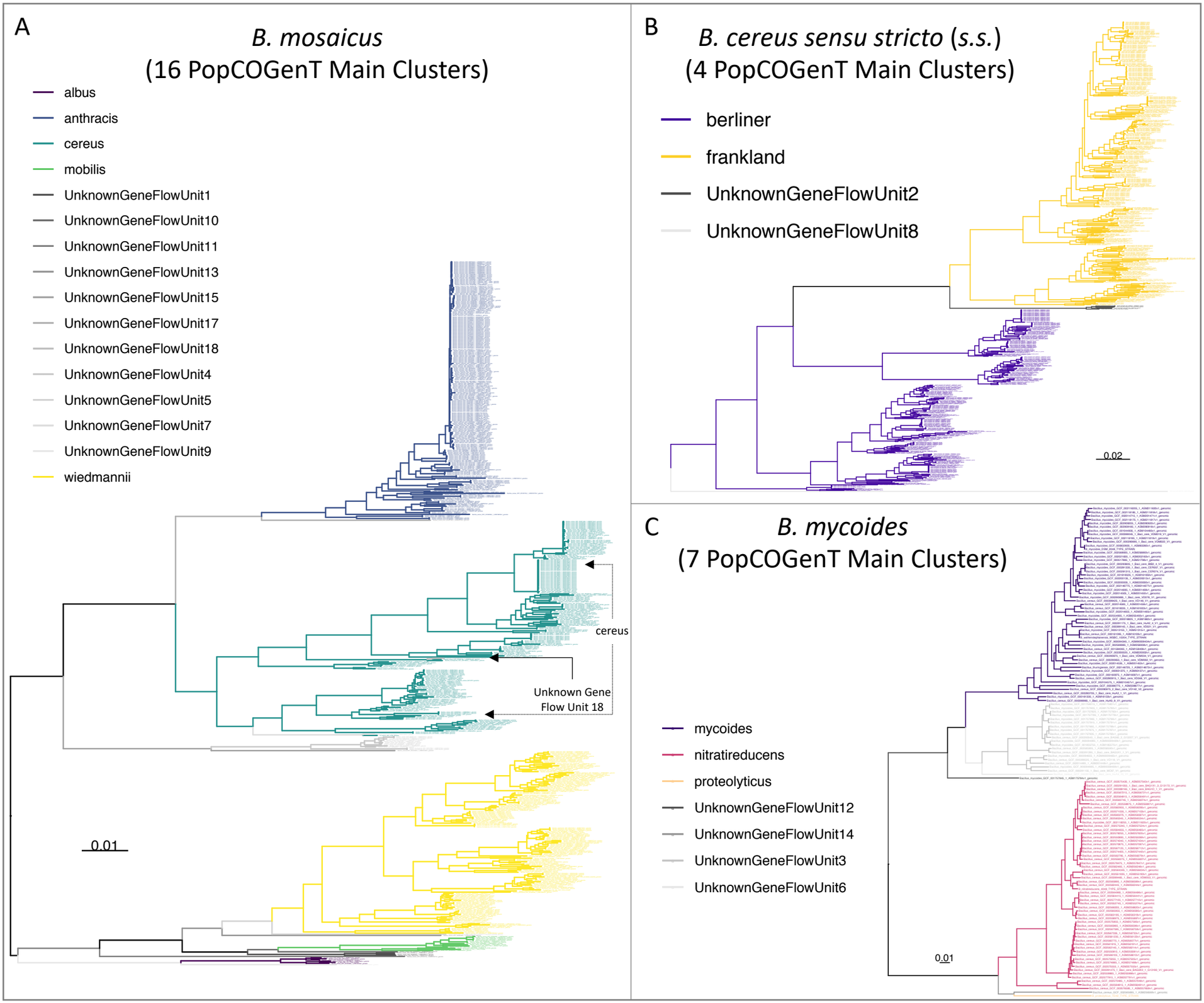
B



GOGO Biological Process Ontology (BPO) Clusters

- ArtQ,HemX,Pit,PstS,YtvI
- ClpX,Def,Gpr,LplJ,YhfN
- CoaD,NuoI,Prs,PyrD,PyrF,PyrK
- DapD,FabI,HipO,HisG,HisH,Icd,IlvA,LeuB,MtnD
- DeoB,MnmA,RapZ,TruB,YfmL,YqeV
- DeoD,GidA,HspX,MutS2,Nfo,PrkA,RuvA
- Der,PetB
- Dxs,FolC,GshB,HemB,PheT,RibBA,YbeY
- EtfA,FadB
- FolD,NuoH,PpaC,YfiH
- FruK,GlcK,SuhB
- GlgA,MraY,MurE,MurG,TagO
- GlpX,PckA
- HepT,MenB
- Hpr,HrcA,YqeH,Zur
- IspA,IspE,PgsA
- KsgA,RluA,RluB,RnmV,RsmE,RsmG,RsuA
- LigT,LpdA,LuxS,Spo0J,YtpP
- LipA,LipL,MobA
- Maf,MinC
- NadA,NadD,PfkA
- SpoIVA,YidC


A
B. toyonensis
B. luti
B. mosaicus
B. paramycoides
B. mycoides
B. pseudomycoides
B. cytotoxicus
B. cereus s.s.
Genomospecies Assignment at 92.5 ANI
B
Cereulide Synthetase Genes cesABCD
C
Sphingomyelinase C (Sph) Gene sph
D
Non-Hemolytic Enterotoxin (Nhe) Genes nheABC
E
Hemolysin BL (Hbl) Genes hblABCD
F
Cytotoxin K (CytK) Variant cytK-1
G
Cytotoxin K (CytK) Variant cytK-2
H
Anthrax Toxin Genes cya, lef, pagA
I
Capsule biosynthesis (Cap) Genes capABCDE
J
Exo-polysaccharide operon (Bps) Genes bpsX-H
K
Hyaluronic acid capsule (Has) Genes hasABC


A
B. mosaicus
(16 PopCOGenT Main Clusters)
albus
anthracis
cereus
mobilis
UnknownGeneFlowUnit1
UnknownGeneFlowUnit10
UnknownGeneFlowUnit11
UnknownGeneFlowUnit13
UnknownGeneFlowUnit15
UnknownGeneFlowUnit17
UnknownGeneFlowUnit18
UnknownGeneFlowUnit4
UnknownGeneFlowUnit5
UnknownGeneFlowUnit7
UnknownGeneFlowUnit9
wiedmannii
cereus
Unknown Gene
Flow Unit 18
0.01
B
B. cereus sensu stricto (s.s.)
(4 PopCOGenT Main Clusters)
berliner
frankland
UnknownGeneFlowUnit2
UnknownGeneFlowUnit8
0.02
C
B. mycoides
(7 PopCOGenT Main Clusters)
mycoides
nitratireducens
proteolyticus
UnknownGeneFlowUnit12
UnknownGeneFlowUnit14
UnknownGeneFlowUnit3
UnknownGeneFlowUnit6
0.01