

Comparison of the performance of multiple whole-genome sequence-based tools for the identification of *Bacillus cereus sensu stricto* biovar *Thuringiensis*

Taejung Chung,¹ Abimel Salazar,¹ Grant Harm,¹ Sophia Johler,² Laura M. Carroll,^{3,4,5,6} Jasna Kovac¹

AUTHOR AFFILIATIONS See affiliation list on p. 9.

ABSTRACT The *Bacillus cereus sensu stricto* (*s.s.*) species comprises strains of biovar *Thuringiensis* (*Bt*) known for their bioinsecticidal activity, as well as strains with foodborne pathogenic potential. *Bt* strains are identified (i) based on the production of insecticidal crystal proteins, also known as Bt toxins, or (ii) based on the presence of *cry*, *cyt*, and *vip* genes, which encode Bt toxins. Multiple bioinformatics tools have been developed for the detection of crystal protein-encoding genes based on whole-genome sequencing (WGS) data. However, the performance of these tools is yet to be evaluated using phenotypic data. Thus, the goal of this study was to assess the performance of four bioinformatics tools for the detection of crystal protein-encoding genes. The accuracy of sequence-based identification of *Bt* was determined in reference to phenotypic microscope-based screening for the production of crystal proteins. A total of 58 diverse *B. cereus sensu lato* strains isolated from clinical, food, environmental, and commercial biopesticide products underwent WGS. Isolates were examined for crystal protein production using phase contrast microscopy. Crystal protein-encoding genes were detected using BtToxin_Digger, BTyper3, IDOPS (identification of pesticidal sequences), and Cry_processor. Out of 58 isolates, the phenotypic production of crystal proteins was confirmed for 18 isolates. Specificity and sensitivity of *Bt* identification based on sequences were 0.85 and 0.94 for BtToxin_Digger, 0.97 and 0.89 for BTyper3, 0.95 and 0.94 for IDOPS, and 0.88 and 1.00 for Cry_processor, respectively. Cry_processor predicted crystal protein production with the highest specificity, and BtToxin_Digger and IDOPS predicted crystal protein production with the highest sensitivity. Three out of four tested bioinformatics tools performed well overall, with IDOPS achieving high sensitivity and specificity (>0.90).

IMPORTANCE Strains of *Bacillus cereus sensu stricto* (*s.s.*) biovar *Thuringiensis* (*Bt*) are used as organic biopesticides. *Bt* is differentiated from the foodborne pathogen *Bacillus cereus s.s.* by the production of insecticidal crystal proteins. Thus, reliable genomic identification of biovar *Thuringiensis* is necessary to ensure food safety and facilitate risk assessment. This study assessed the accuracy of whole-genome sequencing (WGS)-based identification of *Bt* compared to phenotypic microscopy-based screening for crystal protein production. Multiple bioinformatics tools were compared to assess their performance in predicting crystal protein production. Among them, identification of pesticidal sequences performed best overall at WGS-based *Bt* identification.

KEYWORDS *Bacillus thuringiensis*, biopesticide, whole-genome sequencing, Bt toxin

The *Bacillus cereus* group, also known as *B. cereus sensu lato* (*s.l.*), is a species complex that comprises strains with the ability to cause human illness, as well as strains that have agriculturally and industrially beneficial phenotypes. Among well-known species in

Editor Sophie Roussel, Anses, Maisons-Alfort
Laboratory for Food Safety, Maisons-Alfort, France

Address correspondence to Jasna Kovac,
jzk303@psu.edu.

The authors declare no conflict of interest.

See the funding table on p. 9.

Received 6 October 2023

Accepted 19 February 2024

Published 12 March 2024

Copyright © 2024 Chung et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

the *B. cereus* group are foodborne pathogen *B. cereus sensu stricto* (here referred to as *Bc*) (1) and entomopathogen *B. thuringiensis* (here referred to as *Bt*), the latter of which is commercially available as a biopesticide for application in organic farming (2).

The insecticidal activity of *Bt* is supported by the production of insecticidal proteins (i.e., *Bt* toxins), including parasporal crystal proteins such as crystal (Cry) and cytolytic (Cyt) toxins, and non-parasporal proteins such as vegetative insecticidal proteins (Vip) (3, 4). *Bt* toxins act against different insect species, including Lepidoptera, Diptera, Coleoptera, and Hymenoptera (3, 5). In 1995, the United States Environmental Protection Agency (US EPA) registered the first *Bt* biopesticide products for use in the US (6). Currently, over 180 *Bt* products are registered under 15 different EPA product code numbers, and they all include strains of at least one of the four *Bt* subspecies (i.e., *kurstaki*, *israelensis*, *aizawai*, and *tenebrionis*) (6). However, despite the long history of agricultural application of *Bt*, the concerns over safety of *Bt* for humans have been raised over the past decade (7–11).

Bt strains have been reported to encode and produce human enterotoxins that are known to contribute to foodborne illness caused by *Bc* (8, 9, 12). These include hemolysin BL (Hbl), non-hemolytic enterotoxin (Nhe), and cytotoxin K (CytK) (13–16). Given that routine diagnostic assays do not differentiate between *Bt* and *Bc*, it is possible that some *Bc*-associated outbreaks may have been caused by *Bt* (17). However, direct evidence for *Bt* causing human illness has not been explicitly established (17, 18).

The high genomic similarity of *Bc* and *Bt* results in their classification into the same genomospecies, even based on the most conservative classification criteria for genomospecies (13). Moreover, these two species cannot be distinguished by typical culture-based detection methods. The only phenotypic trait that is used for differentiating *Bc* and *Bt* is microscopy-based detection of parasporal crystal proteins (i.e., *Bt* toxins), which are responsible for the bioinsecticidal properties of *Bt* strains (19–21). The above-outlined taxonomic classification shortcomings have been addressed in a new proposed taxonomic framework that considers both genomic and phenotypic information for *B. cereus* group species identification (22). Within the proposed taxonomic framework, *Bc* and *Bt* are represented by one genomospecies, *Bacillus cereus sensu stricto*. Furthermore, a biovar *Thuringiensis* has been defined to represent genomes that carry crystal protein-encoding genes (i.e., *Bt* toxin genes), including *cry*, *cyt*, and *vip* genes (2). Reliable detection of *Bt* toxin genes is therefore required for accurate genome-based detection of *B. cereus s.l.* biovar *Thuringiensis* (*Bt*).

One of the main challenges in detecting *Bt* toxin genes is the high variability in their sequences. For example, over 300 variants of *cry* genes have been identified based on the amino acid sequence similarity (23, 24). Thus, multiple whole-genome sequencing (WGS)-based bioinformatics tools have been developed not only to detect known *Bt* toxin gene sequences but also to predict new variants of these genes. These tools include *BtToxin_Digger* (25), IDOPS (identification of pesticidal sequences) (26), *Cry_processor* (27), and *BTyper3* (13). Each tool has been developed using a different approach. For example, *BtToxin_Digger* uses multiple algorithms, including Basic Local Alignment Search Tool (BLAST), hidden Markov models (HMMs), and support vector machine (SVM) method, to comprehensively detect known and potentially novel *Bt* toxin gene variants. Among these algorithms, BLAST is known to be the most conservative, whereas HMM and SVM model are more likely to produce false positive results and be more suitable for the detection of novel gene variants. IDOPS uses profile HMMs to predict *Bt* toxin-encoding genes and annotates sequences adjacent to predicted *Bt* toxin-encoding genes, which can help users interpret the genetic context in which these genes are detected. *Cry_processor* searches for 3-domain *cry* gene sequences using profile HMMs, which has the limitation of missing other genes (e.g., *cyt* and *vip*). Lastly, *BTyper3* uses BLAST to search for crystal protein-encoding genes based on a specific amino acid similarity threshold and hence may not be able to detect novel variants of *Bt* toxin genes.

Despite the availability of these tools, their performance has not been assessed using phenotypic assays. This study therefore aimed to (i) compare Bt toxin-encoding gene identification tools and (ii) compare their performance against phenotypic data (i.e., microscopy screening for crystal protein production).

RESULTS AND DISCUSSION

We first compared the completeness of hybrid and short-read genome assemblies using the N50 metric and the number of contigs, as determined by QUAST (Fig. 1). N50 values for hybrid assemblies (min = 443,575 bp, max = 5,675,203 bp, mean = 4,245,832.3 bp) were significantly higher than N50 for short-read assemblies (min = 22,040 bp, max = 514,228 bp, mean = 113,641.28 bp) (t -test P value = 1.98×10^{-47}). The number of contigs equal to or longer than 1,000 bp was significantly higher in short-read assemblies (min = 32, max = 607, mean = 235) compared to hybrid assemblies (min = 1, max = 98, mean = 17) (t -test P value = 7.37×10^{-23}). N50 values were, on average, 59% shorter in short-read assemblies than hybrid assemblies. Similarly, there were on average 79% fewer contigs in hybrid assemblies than short-read assemblies. Both of these metrics taken together indicate that hybrid assemblies were more complete. Both types of assemblies were further used for the assessment of bioinformatics tools for the detection of Bt toxin-encoding genes to measure the effect of assembly completeness on the performance of individual tools.

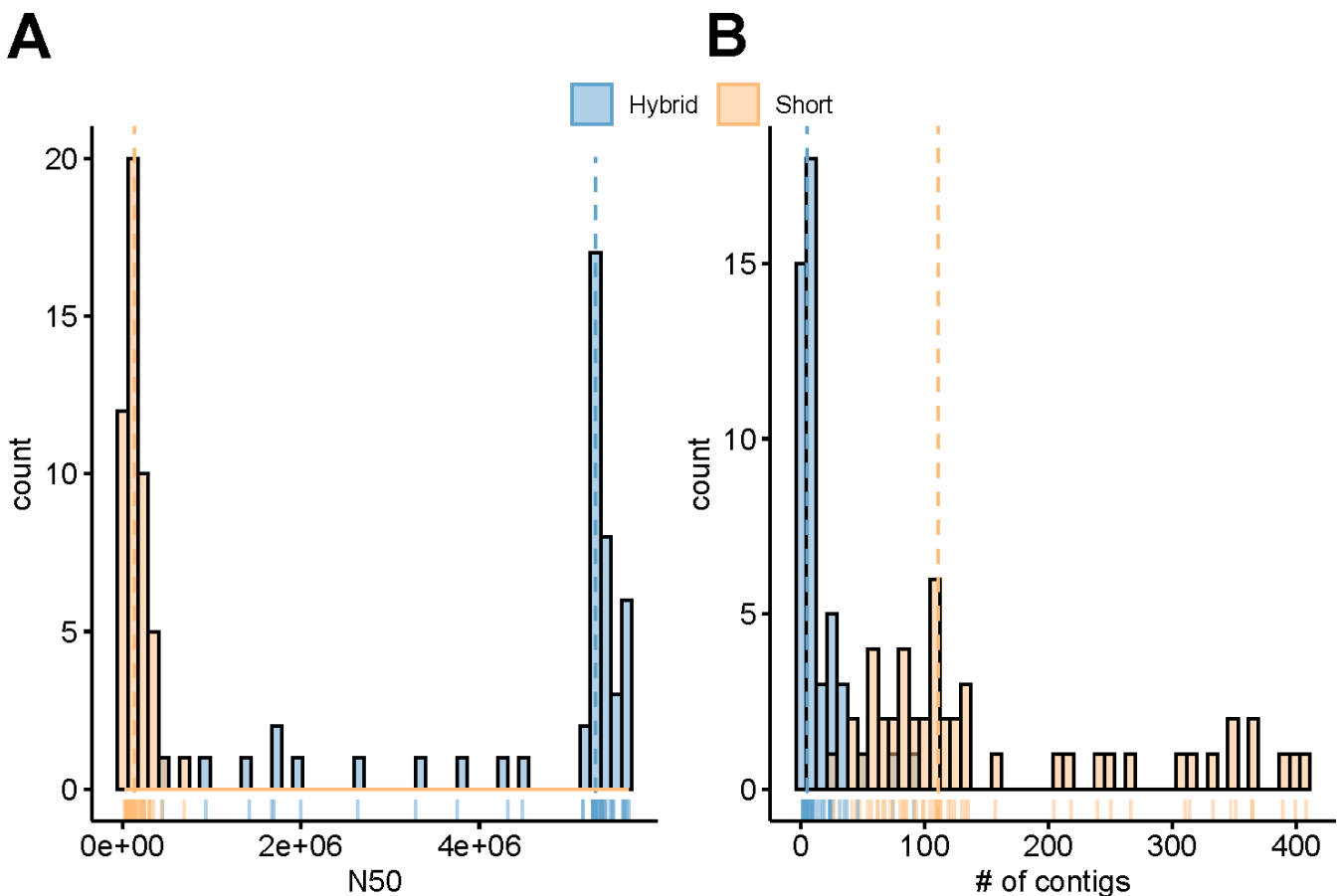


FIG 1 Distribution of contig counts in hybrid and short-read assemblies. (A) N50 and (B) number of total contigs ($\geq 1,000$ bp).

IDOPS performed best overall for Bt toxin-encoding gene detection in comparison to phenotypic data

Prior to the application of bioinformatics tools for the detection of Bt toxin-encoding genes, we identified and removed redundant genomes, here defined as genomes that differed by <6 single nucleotide polymorphisms (SNPs). Briefly, using PubMLST's seven-gene multi-locus sequence typing (MLST) scheme for "*B. cereus*," 37 different sequence types (STs) were identified with biovar *Thuringiensis* identified in 10 STs (i.e., ST1085, ST1099, ST1142, ST138, ST15, ST1734, ST325, ST33, ST414, and ST8). Out of the 37 different MLST STs, 6 STs had more than one isolate (i.e., ST1099, ST1424, ST15, ST8, ST24, and ST73); thus, high-quality SNPs were identified within those STs. As a result, a total of 20 clonal genomes were excluded from further analyses to mitigate clonal redundancy bias. This resulted in a total of 58 isolates that were included in the assessment of the performance of four bioinformatics tools (i.e., IDOPS, Cry_processor, BtToxin_Digger, and BTyper3). IDOPS uses profile HMMs (28) to detect pesticidal protein-encoding genes in accordance with the BPPRC (Bacterial Pesticidal Protein Resource Center) nomenclature system (23). Cry_processor was developed based on an HMM-based algorithm and has two different search modes available: domain only (DO) and find domain (FD). DO mode detects the sequences comprising all the three domains in the right order, and FD mode searches novel domains of the matched Bt toxin gene sequences, using the Bt-Toxin nomenclature database (29). Here, we tested both modes to compare the performance of the pipeline and found no differences in results between the two modes. BtToxin_Digger uses three different methods (i.e., BLAST, HMM, and SVM) for the identification of Bt toxin gene sequences using the Bt toxin nomenclature database. Lastly, BTyper3 uses BLAST to detect Bt toxin genes using translated amino acid sequences with conservative detection thresholds of 70% identity and 50% coverage.

The production of crystal proteins was confirmed for 18 out of 58 isolates using phase contrast microscopy (Fig. 2). When applied to short-read assemblies, CryProcessor predicted crystal protein production with the highest specificity (0.97). BtToxin_Digger and IDOPS predicted crystal protein production with the highest sensitivity (0.94), whereas BtToxin_Digger had the lowest specificity (0.82). When applied to hybrid assemblies, the same bioinformatics programs showed the highest and lowest specificities and sensitivities (Table 1). Specifically, CryProcessor predicted protein production with higher specificity (1.00) compared to the other three programs: BTyper3 (0.97), BtToxin_Digger (0.85), and IDOPS (0.95). However, the other three programs showed higher sensitivity than CryProcessor (0.83): BTyper3 (0.88), BtToxin_Digger (0.94), and IDOPS (0.94). Similar to results observed for short-read assemblies, BtToxin_Digger had the lowest specificity when applied to hybrid assemblies (0.82). These results indicate that genome assembly completeness was not a major factor influencing the performance of bioinformatics tools for the detection of Bt toxin-encoding genes. Three out of four tested bioinformatics tools (i.e., IDOPS, BTyper3, and Cry_processor) performed well

TABLE 1 Sensitivity, specificity, and positive and negative predictive values for predicting crystal protein production using four bioinformatics tools

Assembly	Programs	Sensitivity	Specificity	PPV ^a	NPV ^b
Hybrid	BTyper3	0.89	0.97	0.94	0.95
	BtToxin_Digger	0.94	0.85	0.73	0.97
	CryProcessor	0.88	1.00	1	0.95
	IDOPS ^c	0.94	0.95	0.89	0.97
Short read	BTyper3	0.88	0.95	0.88	0.95
	BtToxin_Digger	0.94	0.82	0.70	0.97
	CryProcessor	0.83	0.97	0.93	0.93
	IDOPS ^c	0.94	0.92	0.85	0.97

^aPPV, positive predictive value.

^bNPV, negative predictive value.

^cIDOPS, identification of pesticidal sequences.

isolates were not phylogenetically closely related and were predicted as positive for Bt toxin-encoding genes by IDOPS' profile HMMs 5 (CryM5) and 6 (CryM6). All other IDOPS models (i.e., CryM1 to M4, CytM1, CytM2, and VipM1) correctly predicted the presence of Bt toxin-encoding genes. The CryM5 profile HMM targets Cry proteins with less conserved variants of the classical 3 domains, while CryM6 targets the C-terminal region of Cry pesticidal proteins (26). This could explain why the CryM5 and CryM6 models produced false positive predictions on non-*Bt* strains.

BtToxin_Digger, which had the lowest specificity, uses a combination of multiple algorithms, including BLAST, HMM, and SVM. We examined whether false positive results were caused by any specific algorithm and found that the majority of false positive calls were attributable to the outcomes of predictive models (i.e., HMM and SVM) (Table 2). HMM had both low sensitivity and specificity ($S_n = 0.68$, $S_p = 0.68$), whereas SVM had low sensitivity and high specificity ($S_n = 0.00$, $S_p = 0.95$). In contrast to HMM and SVM, we found that BLAST search alone performed very well ($S_n = 0.97$, $S_p = 0.95$) and produced BtToxin_Digger results that were comparable to the other three programs. Although the HMM and SVM model implemented in BtToxin_Digger produce more false positive results compared to the BLAST approach, they may be more useful for the discovery of novel insecticidal protein-encoding genes compared to BLAST or other bioinformatics tools tested here.

Phylogenetic distribution of biovar *Thuringiensis* strains within the *B. cereus* group

Overall, 12 out of 18 biovar *Thuringiensis* strains were classified into adjusted eight-group *panC* phylogenetic group IV, 2 into group II, 1 into group V, and 3 into group VI (Fig. 2). The maximum-likelihood phylogenetic analysis showed that the biovar *Thuringiensis* strains (as defined by the detection of Bt toxin-encoding genes by at least three bioinformatics tools or the microscopic screening) were found across different lineages of phylogenetic group IV (corresponding to *B. cereus* s.s.) and were intermixed with non-*Thuringiensis* strains (Fig. 2). Notably, most of the biovar *Thuringiensis* strains included in this study were not closely related to the *B. thuringiensis* type strain (ATCC 10792) nor the *B. cereus* s.s. type strain (ATCC 14579). An average SNP difference relative to the *B. thuringiensis* type strain was 44,765 SNPs (min = 7,717 SNPs, max = 76,229 SNPs). Compared to the *B. cereus* s.s. type strain, tested strains differed by an average of 37,748.94 SNPs (min = 6,808 SNPs, max = 62,696 SNPs). This demonstrates that phylogenetic analysis alone is not sufficient for the identification of strains belonging to biovar *Thuringiensis*. This finding agrees with previous studies and suggests that Bt toxin genes (e.g., *cry*, *cyt*, and *vip*) likely disseminated among *B. cereus* group strains through horizontal gene transfer (16).

Bt toxin genes are typically found on plasmids (30, 31), although they have also been detected within the chromosome of the *B. thuringiensis* HER1410 strain (32). We found Bt toxin-encoding genes on non-chromosomal contigs in 17 out of 18 hybrid assemblies of strains belonging to biovar *Thuringiensis*. Only one isolate (PS00095) carried the *cry27Aa1* gene on a chromosomal contig. This gene shared 75% identity and 100% coverage with the reference sequence in the *B. thuringiensis* delta-endotoxin nomenclature database. A transfer and chromosome integration of a plasmid-borne gene sequence may have been

TABLE 2 Sensitivity, specificity, and positive and negative predictive values for predicting crystal protein production using different BtToxin_Digger algorithms (i.e., BLAST, HMM, and SVM model) applied on hybrid assemblies

Algorithm ^a	Sensitivity	Specificity	PPV ^b	NPV ^b
BLASTP	0.97	0.95	0.94	0.98
HMM	0.68	0.68	0.62	0.73
SVM	0.00	0.95	0.00	0.55

^aBLASTP, Basic Local Alignment Search Tool for protein sequences.

^bPPV, positive predictive value; NPV, negative predictive value.

facilitated by a transposase; however, no transposable elements were detected adjacent to *cry27Aa1*.

Overall, this work demonstrates that sequence-based detection of *cry*, *cyt*, and *vip* can serve as an accurate method for the detection of *B. cereus* biovar *Thuringiensis*, which can aid in the identification of this biovar, as defined in the taxonomic nomenclature proposed by Carroll et al. (13) or a variation of it agreed upon by the stakeholders.

MATERIALS AND METHODS

Isolates included in the study

A total of 78 *B. cereus* group strains available in the Kovac lab culture collection, some of which have been reported previously (15, 33), were included in the study based on the following criteria: (i) an isolate was classified as *B. cereus* s.s. (phylogenetic group IV) using BTyper 3 (v.3.2.0) ($n = 72$) or (ii) an isolate belonged to any of the *B. cereus* s.l. phylogenetic groups and had *cry*, *cyt*, and/or *vip* genes detected in its draft genome using BTyper3 ($n = 6$) (v.3.2.0) (13). Isolate information, including year of isolation, origin of isolation, and NCBI accession numbers are listed in the Supplemental Material (Table S1).

Illumina whole-genome sequencing and sequence quality control

Total genomic DNA was extracted from isolates using E.Z.N.A Bacterial DNA extraction kit (Omega Bio-Tek, Georgia, USA) by following the manufacturer's instructions. Extracted DNA was examined for quality and quantity using Nanodrop One (Thermo Fisher Scientific, Massachusetts, USA) and Qubit 3 (Thermo Fisher Scientific, Massachusetts, USA), respectively. DNA libraries were prepared using a NexteraXT library preparation kit (Illumina, California, USA). Samples were sequenced using an Illumina NextSeq with 150-bp paired-end reads. Additionally, genomes of 15 isolates with publicly available WGS reads were obtained from the NCBI Sequence Read Archive (SRA; NCBI BioProject accessions [PRJNA437714](https://www.ncbi.nlm.nih.gov/PRJNA437714) and [PRJNA288462](https://www.ncbi.nlm.nih.gov/PRJNA288462)) (15, 33). Illumina adapters and low-quality reads were trimmed using Trimmomatic (v.0.39) with a sliding window size of 4 and quality cutoff value of 15 (SLIDINGWINDOW:4:15) (34). Reads shorter than 36 bp were excluded (MINLEN:36). Trimmed reads qualities were assessed using FastQC (v.0.11.9) (35).

Nanopore whole-genome sequencing and sequence quality control

DNA was extracted using the QIAamp DNA blood mini kit (Qiagen, Hilden, Germany) by following the manufacturer's instructions with additional steps for cell lysis. Briefly, 3 loopfuls of biomass grown at 30°C on brain-heart infusion agar (BD Biosciences, New Jersey, USA) was collected and mixed with a phosphate-buffered saline (PBS) buffer (137 mM NaCl, 1.7 mM KCl, 10 mM Na₂HPO₄, 1.8 mM KH₂PO₄). One gram of 0.1-mm zirconia/silica beads (Biospec Products, Oklahoma, USA) was added to tubes, which were then vortexed in a horizontal position to disrupt cell walls. The Oxford Nanopore Technologies (ONT) RBK-004 rapid barcoding kit was used for the library preparation by following the manufacturer's instructions, and four libraries were pooled for sequencing (ONT, Oxford, UK). An R9 flow cell (FLO-MIN106; ONT, Oxford, UK) was used for sequencing using a MinION Mk1C device (ONT). Raw sequencing signal files were used for high-accuracy basecalling and adapter trimming using Guppy (v.6.0.1). Low-quality reads were trimmed using FiltLong (v.0.2.1) (36), and the quality of trimmed reads was assessed using FastQC (v.0.11.9).

Genome assembly

Illumina reads were assembled *de novo* with SPAdes using the "isolate" mode (--isolate) and *k*-mer sizes of 99 or 127 (-k 99,127) (v.3.15.3) (37). Trimmed Illumina short reads and

trimmed ONT long reads were used for hybrid genome assembly with Unicycler using default parameters (v.0.5.0) (38). Assembly quality was assessed using QUAST (v.5.0.2) (39).

Maximum-likelihood phylogenetic tree construction

To visualize phylogenetic relationships among isolates, core-genome single nucleotide polymorphisms (cgSNPs) of assembled genomes were identified by kSNP3 (v.3.1.2) (40), using $k = 21$; assembled genomes were used as input, along with the *B. cereus* ATCC 14579 and *B. thuringiensis* ATCC 10792 type strain genomes (GenBank accession numbers [GCA_018309165](#) and [GCA_000161615](#), respectively). Maximum-likelihood (ML) phylogenetic trees were constructed using the detected cgSNPs and IQ-TREE 2 (v.2.2.0), using the generalized time-reversible model with a gamma distribution and ascertainment bias correction (GTR + G + ASC), plus 1,000 ultrafast bootstraps (41–44). The final ML tree was visualized and annotated using iTOL (v.6.8) (45).

To remove redundant and clonal isolates, high-quality SNPs were identified using the FDA CFSAN SNP pipeline with default setting (v.2.2.1) (46). This analysis was conducted separately for isolates belonging to each individual MLST sequence type that contained two or more isolates. MLST STs were assigned using BTyper 3 (v.3.2.0) and PubMLST's seven-gene MLST scheme for "*B. cereus*" (47). Pairwise SNP differences were calculated within each ST, and for each group of clonal genomes (i.e., isolates that differed by <6 whole-genome SNPs), one high-quality representative genome was selected based on N50 values and number of contigs. The same SNP pipeline was applied to all identified biovar *Thuringiensis* genomes by using two type strains as references (*B. thuringiensis* type strain ATCC 10792 and *B. cereus* type strain ATCC 14579) to calculate average SNP differences between biovar *Thuringiensis* strains and the two type strains.

Bioinformatic detection of Bt toxin-encoding genes

Bt toxin-encoding genes were detected in each assembled genome using each of the following bioinformatic pipelines: BTyper3 (v.3.2.0) (13), BtToxin_Digger (v.1.0.10) (25), IDOPS (v.0.2.2) (26), and Cry_processor (last update on August 2019; DO and FD were used) (27) (Table 3). The four bioinformatics tools were applied to both short-read and hybrid assemblies. Protein coding sequences (CDSs) were identified using Prokka with default setting (v.1.14.5) (48). The resulting CDSs were used as an input for IDOPS, Cry_processor, and BtToxin_Digger. Nucleotide genome assemblies were used as an input for BTyper3.

Microscopy-based screening of isolates for the production of insecticidal crystal proteins

Isolates were grown on T3 agar for 3 days at 30°C to promote sporulation and crystal protein production. After completed incubation, one colony was resuspended in 10 μ L of PBS on a microscope slide and covered with a cover slip (VWR international, Pennsylvania, USA) and screened within 10 minutes for the presence of crystal proteins using a phase-contrast microscope (Olympus BX51, Olympus, Tokyo, Japan). Crystal proteins were differentiated by shape (bipyramidal, cuboidal, and circular), size (smaller than a vegetative cell and endospore of *Bacillus*), and color (darker than *Bacillus* endospore). Initial screening results were classified as "positive," "negative," or "inconclusive."

TABLE 3 Description of bioinformatics tools for prediction of Bt toxin-encoding genes

Program	Method ^a	Reference database	Reference
BTyper3	BLASTP	The <i>Bacillus thuringiensis</i> delta-endotoxin nomenclature	(13)
BtToxin_Digger	BLASTP, HMM, SVM	The <i>Bacillus thuringiensis</i> delta-endotoxin nomenclature	(25)
CryProcessor	HMM (diamond)	The <i>Bacillus thuringiensis</i> delta-endotoxin nomenclature, NCBI SRA, IPG database	(27)
IDOPS	Profile HMM	The BPPRC	(26)

^aBLASTP, Basic Local Alignment Search Tool for protein sequences.

Screening was repeated three times by two individuals to assess the production of crystal proteins.

Statistical analysis

All statistical analyses in the study were conducted using R (v4.2.1) (49). A paired *t*-test was employed to compare the average N50 and the number of contigs between short-read assembly and hybrid assembly (50). The sensitivity, specificity, positive predictive value, and negative predictive value were calculated to assess the performance of four different bioinformatics tools for the detection of Bt toxin-encoding genes (51). Genome assemblies produced using short reads as well as hybrid assemblies were used. The results of microscopy screening were used to validate the sequence-based bioinformatic prediction of *B. cereus* biovar Thuringiensis.

ACKNOWLEDGMENTS

This work was supported by the USDA National Institute of Food and Agriculture and Hatch Appropriations under project PEN04853 and accession 7005519, multistate project 4666, and USDA Agriculture and Food Research Initiative Research and Extension Experiences for Undergraduates grant 2021-67037-34628 “Engaging Students from Undergraduate-Centric Institutions in Research, Informatics, and Experiential Opportunities in Food Microbiology.” L.M.C. was supported by the SciLifeLab & Wallenberg Data Driven Life Science Program (grant: KAW 2020.0239).

AUTHOR AFFILIATIONS

¹Department of Food Science, The Pennsylvania State University, University Park, Pennsylvania, USA

²Institute for Food Safety and Hygiene, Vetsuisse Faculty, University of Zurich, Zurich, Switzerland

³Department of Clinical Microbiology, SciLifeLab, Umeå University, Umeå, Sweden

⁴Laboratory for Molecular Infection Medicine Sweden (MIMS), Umeå University, Umeå, Sweden

⁵Umeå Centre for Microbial Research (UCMR), Umeå University, Umeå, Sweden

⁶Integrated Science Lab (IceLab), Umeå University, Umeå, Sweden

AUTHOR ORCID*s*

Taejung Chung  <http://orcid.org/0000-0003-4783-6473>

Sophia Johler  <http://orcid.org/0000-0003-4299-5651>

Jasna Kovac  <http://orcid.org/0000-0002-9465-4552>

FUNDING

Funder	Grant(s)	Author(s)
USDA National Institute of Food and Agriculture	PEN04853 accession 7005519, 4666, 2021-67037-34628	Jasna Kovac
SciLifeLab & Wallenberg Data Driven Life Science Program	KAW 2020.0239	Laura M. Carroll

DATA AVAILABILITY

Paired-end Illumina reads sequenced in this study have been deposited in the NCBI SRA database under BioProject accession number [PRJNA715191](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA715191). Single-end Nanopore sequences have been deposited in the NCBI SRA database under BioProject accession number [PRJNA1010762](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1010762). Detailed information for sequencing data used in this study is available in Table S1. Scripts for bioinformatic analyses and results obtained using each

bioinformatics tool are available in the GitHub repository (https://github.com/tuc289/Bt_toxin_tools_validation).

ADDITIONAL FILES

The following material is available [online](#).

Supplemental Material

Table S1 (AEM01778-23-s0001.docx). Name, phylogenetic groups, sources, and NCBI SRA accessions for isolates used in this study.

REFERENCES

1. Granum PE, Lund T. 1997. *Bacillus cereus* and its food poisoning toxins. FEMS Microbiol Lett 157:223–228. <https://doi.org/10.1111/j.1574-6968.1997.tb12776.x>
2. Bravo A, Likitvatanavong S, Gill SS, Soberón M. 2011. *Bacillus thuringiensis*: a story of a successful bioinsecticide. Insect Biochem Mol Biol 41:423–431. <https://doi.org/10.1016/j.ibmb.2011.02.006>
3. Palma L, Muñoz D, Berry C, Murillo J, Caballero P. 2014. *Bacillus thuringiensis* toxins: an overview of their biocidal activity. Toxins (Basel) 6:3296–3325. <https://doi.org/10.3390/toxins6123296>
4. Estruch JJ, Warren GW, Mullins MA, Nye GJ, Craig JA, Koziel MG. 1996. Vip3A, a novel *Bacillus thuringiensis* vegetative insecticidal protein with a wide spectrum of activities against lepidopteran insects. Proc Natl Acad Sci U S A 93:5389–5394. <https://doi.org/10.1073/pnas.93.11.5389>
5. Liu X, Ruan L, Peng D, Li L, Sun M, Yu Z. 2014. Thuringiensin: a thermostable secondary metabolite from *Bacillus thuringiensis* with insecticidal activity against a wide range of insects. Toxins (Basel) 6:2229–2238. <https://doi.org/10.3390/toxins6082229>
6. EPA. 1998. Reregistration eligibility decision (RED) - *Bacillus thuringiensis*. https://www3.epa.gov/pesticides/chem_search/reg_actions/reregistration/red_PC-006400_30-Mar-98.pdf
7. Biggel M, Etter D, Corti S, Brodmann P, Stephan R, Ehling-Schulz M, Johler S. 2021. Whole genome sequencing reveals biopesticidal origin of *Bacillus thuringiensis* in foods. Front Microbiol 12:775669. <https://doi.org/10.3389/fmicb.2021.775669>
8. Damgaard PH, Larsen HD, Hansen BM, Bresciani J, Jørgensen K. 1996. Enterotoxin-producing strains of *Bacillus thuringiensis* isolated from food. Lett Appl Microbiol 23:146–150. <https://doi.org/10.1111/j.1472-765x.1996.tb00051.x>
9. Gaviña Rivera AM, Granum PE, Priest FG. 2000. Common occurrence of enterotoxin genes and enterotoxicity in *Bacillus thuringiensis*. FEMS Microbiol Lett 190:151–155. <https://doi.org/10.1111/j.1574-6968.2000.tb09278.x>
10. Johler S, Kalbhenn EM, Heini N, Brodmann P, Gautsch S, Bağcıoğlu M, Contzen M, Stephan R, Ehling-Schulz M. 2018. Enterotoxin production of *Bacillus thuringiensis* isolates from biopesticides, foods, and outbreaks. Front Microbiol 9:1915. <https://doi.org/10.3389/fmicb.2018.01915>
11. Schwenk V, Riegg J, Lacroix M, Märklbauer E, Jessberger N. 2020. Enteropathogenic potential of *Bacillus thuringiensis* isolates from soil, animals, food and biopesticides. Foods 9:1484. <https://doi.org/10.3390/foods9101484>
12. Griffiths MW. 1990. Toxin production by psychrotrophic *Bacillus* spp. present in milk. J Food Prot 53:790–792. <https://doi.org/10.4315/0362-028X-53.9.790>
13. Carroll LM, Cheng RA, Kovac J. 2020. No assembly required: using BType3 to assess the congruency of a proposed taxonomic framework for the *Bacillus cereus* group with historical typing methods. Front Microbiol 11:580691. <https://doi.org/10.3389/fmicb.2020.580691>
14. Kovac J, Miller RA, Carroll LM, Kent DJ, Jian J, Beno SM, Wiedmann M. 2016. Production of hemolysin BL by *Bacillus cereus* group isolates of dairy origin is associated with whole-genome phylogenetic clade. BMC Genomics 17:581. <https://doi.org/10.1186/s12864-016-2883-z>
15. Carroll LM, Wiedmann M, Mukherjee M, Nicholas DC, Mingle LA, Dumas NB, Cole JA, Kovac J. 2019. Characterization of emetic and diarrheal *Bacillus cereus* strains from a 2016 foodborne outbreak using whole-genome sequencing: addressing the microbiological, epidemiological, and bioinformatic challenges. Front Microbiol 10:144. <https://doi.org/10.3389/fmicb.2019.00144>
16. Biggel M, Jessberger N, Kovac J, Johler S. 2022. Recent paradigm shifts in the perception of the role of *Bacillus thuringiensis* in foodborne disease. Food Microbiol 105:104025. <https://doi.org/10.1016/j.fm.2022.104025>
17. Bonis M, Felten A, Pairaud S, Dijoux A, Maladen V, Mallet L, Radomski N, Duboisset A, Arar C, Sarda X, Vial G, Mistou M-Y, Firmesse O, Hennekinne J-A, Herbin S. 2021. Comparative phenotypic, genotypic, and genomic analyses of *Bacillus thuringiensis* associated with foodborne outbreaks in France. PLoS One 16:e0246885. <https://doi.org/10.1371/journal.pone.0246885>
18. Raymond B, Federici BA. 2017. In defence of *Bacillus thuringiensis*, the safest and most successful microbial insecticide available to humanity—a response to EFSA. FEMS Microbiol Ecol 93:fix084. <https://doi.org/10.1093/femsec/fix084>
19. Carroll LM, Cheng RA, Wiedmann M, Kovac J. 2022. Keeping up with the *Bacillus cereus* group: taxonomy through the genomics era and beyond. Crit Rev Food Sci Nutr 62:7677–7702. <https://doi.org/10.1080/10408398.2021.1916735>
20. Ehling-Schulz M, Messelhäusser U. 2013. *Bacillus* “next generation” diagnostics: moving from detection toward subtyping and risk-related strain profiling. Front Microbiol 4:32. <https://doi.org/10.3389/fmicb.2013.00032>
21. Tallent SM, Rhodehamel EJ, Harmon SM, Bennett RW. 1998. BAM: *Bacillus cereus*. In Bacteriological analytical manual, 8th ed. US Food and Drug Administration, Silver Spring.
22. Carroll LM, Wiedmann M, Kovac J. 2020. Proposal of a taxonomic nomenclature for the *Bacillus cereus* group which reconciles genomic definitions of bacterial species with clinical and industrial phenotypes. mBio 11:e00034-20. <https://doi.org/10.1128/mBio.00034-20>
23. Panneerselvam S, Mishra R, Berry C, Crickmore N, Bonning BC. 2022. BPPRC database: a web-based tool to access and analyse bacterial pesticidal proteins. Database (Oxford) 2022:baac022. <https://doi.org/10.1093/database/baac022>
24. Crickmore N, Berry C, Panneerselvam S, Mishra R, Connor TR, Bonning BC. 2021. A structure-based nomenclature for *Bacillus thuringiensis* and other bacteria-derived pesticidal proteins. J Invertebr Pathol 186:107438. <https://doi.org/10.1016/j.jip.2020.107438>
25. Liu H, Zheng J, Bo D, Yu Y, Ye W, Peng D, Sun M. 2021. BtToxin_Digger: a comprehensive and high-throughput pipeline for mining toxin protein genes from *Bacillus thuringiensis*. Bioinformatics 38:250–251. <https://doi.org/10.1093/bioinformatics/btab506>
26. Díaz-Valerio S, Lev Hacoheh A, Schöppe R, Liesegang H. 2021. IDOPS, a profile HMM-based tool to detect pesticidal sequences and compare their genetic context. Front Microbiol 12:664476. <https://doi.org/10.3389/fmicb.2021.664476>
27. Shikov AE, Malovichko YV, Skitchenko RK, Nizhnikov AA, Antonets KS. 2020. No more tears: mining sequencing data for novel *Bt* Cry toxins with CryProcessor. Toxins (Basel) 12:204. <https://doi.org/10.3390/toxins12030204>
28. Eddy SR. 1998. Profile hidden markov models. Bioinformatics 14:755–763. <https://doi.org/10.1093/bioinformatics/14.9.755>
29. Crickmore N, Zeigler D, Schnepf E, VanJ, Lereclus D, Baum J, Bravo A, Dean D. 2016. *Bacillus thuringiensis* toxin nomenclature. Available from: <http://www.btomenclature.info>

30. Gillis A, Guo S, Bolotin A, Makart L, Sorokin A, Mahillon J. 2017. Detection of the cryptic prophage-like molecule *pBtic235* in *Bacillus thuringiensis* subsp. *israelensis*. *Res Microbiol* 168:319–330. <https://doi.org/10.1016/j.resmic.2016.10.004>
31. Fiedoruk K, Daniluk T, Mahillon J, Leszczynska K, Swiecicka I. 2017. Genetic environment of *cry1* genes indicates their common origin. *Genome Biol Evol* 9:2265–2275. <https://doi.org/10.1093/gbe/evx165>
32. Lechuga A, Lood C, Salas M, van Noort V, Lavigne R, Redrejo-Rodríguez M. 2020. Completed genomic sequence of *Bacillus thuringiensis* HER1410 reveals a *Cry*-containing chromosome, two megaplasmids, and an integrative plasmidial prophage. *G3 (Bethesda)* 10:2927–2939. <https://doi.org/10.1534/g3.120.401361>
33. Miller RA, Jian J, Beno SM, Wiedmann M, Kovac J. 2018. Intraclade variability in toxin production and cytotoxicity of *Bacillus cereus* group type strains and dairy-associated isolates. *Appl Environ Microbiol* 84:e02479-17. <https://doi.org/10.1128/AEM.02479-17>
34. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30:2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>
35. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data
36. Wick RR, Menzel P. 2018. Filtlong. Available from: github.com/rrwick/Filtlong. Retrieved 15 Aug 2021.
37. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455–477. <https://doi.org/10.1089/cmb.2012.0021>
38. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 13:e1005595. <https://doi.org/10.1371/journal.pcbi.1005595>
39. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUASt: quality assessment tool for genome assemblies. *Bioinformatics* 29:1072–1075. <https://doi.org/10.1093/bioinformatics/btt086>
40. Gardner SN, Slezak T, Hall BG. 2015. kSNP3. 0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics* 31:2877–2878. <https://doi.org/10.1093/bioinformatics/btv271>
41. Lewis PO. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst Biol* 50:913–925. <https://doi.org/10.1080/106351501753462876>
42. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 32:268–274. <https://doi.org/10.1093/molbev/msu300>
43. Tarvaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. In *Some mathematical question in biology-DNA sequence analysis*
44. Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* 39:306–314. <https://doi.org/10.1007/BF00160154>
45. Letunic I, Bork P. 2021. Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* 49:W293–W296. <https://doi.org/10.1093/nar/gkab301>
46. Davis S, Pettengill JB, Luo Y, Payne J, Shpuntoff A, Rand H, Strain E. 2015. CFSAN SNP pipeline: an automated method for constructing SNP matrices from next-generation sequence data. *PeerJ Comput Sci* 1:e20. <https://doi.org/10.7717/peerj-cs.20>
47. Jolley KA, Bray JE, Maiden MCJ. 2018. Open-access bacterial population genomics: BIGSdb software, the PubMLST. org website and their applications. *Wellcome Open Res* 3:124. <https://doi.org/10.12688/wellcomeopenres.14826.1>
48. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068–2069. <https://doi.org/10.1093/bioinformatics/btu153>
49. R Core Team. 2020. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org>.
50. Student. 1908. The probable error of a mean. *Biometrika* 6:1. <https://doi.org/10.2307/2331554>
51. Altman DG, Bland JM. 1994. Diagnostic tests. 1: sensitivity and specificity. *BMJ* 308:1552. <https://doi.org/10.1136/bmj.308.6943.1552>