Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Species identity does not predict pesticidal gene content in the Bacillus thuringiensis group: comparative genomics of strains originally recognized as B. thuringiensis against reference genomes

Cristian A. Flores-Cruz
Universidad Autónoma de Nuevo León
Ezequiel Lozano-Lieberenz
Universidad Autónoma de Nuevo León
Andre Segura-Ruiz
Universidad Autónoma de Nuevo León
Yvonne He
Memorial University of Newfoundland, Newfoundland and Labrador
Andrew S. Lang
Memorial University of Newfoundland, Newfoundland and Labrador
José Luis Galán-González
Universidad Autónoma de Nuevo León
Saúl A. Martínez-Morales
Universidad Autónoma de Nuevo León
Myriam Elías Santos
Universidad Autónoma de Nuevo León
Luis J. Galán-Wong
Universidad Autónoma de Nuevo León
Benito Pereyra-Alférez
bpereyra@gmail.com
Universidad Autónoma de Nuevo León

Research Article

Keywords: Bacillus cereus group, Bacillus thuringiensis, average nucleotide identity, biological control, comparative genomics, insertion sequences, mobile genetic elements, pesticidal proteins, plasmids.

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11483591/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Title

Species identity does not predict pesticidal gene content in the Bacillus thuringiensis group: comparative genomics of strains originally recognized as B. thuringiensis against reference genomes

Authors and affiliations

Cristian A. Flores-Cruz[1,2], Ezequiel Lozano-Lieberenz[1], Andre Segura-Ruiz[1], Yvonne He[2], Andrew S. Lang[2], José Luis Galán-González[1], Saúl A. Martínez-Morales[1], Myriam Elías Santos[1], Luis J. Galán-Wong[1], Benito Pereyra-Alférez[1,*]

[1]Universidad Autónoma de Nuevo León, Facultad de Ciencias Biológicas, Instituto de Biotecnología, San Nicolás de los Garza, Nuevo León, 66450 México

[2]Department of Biology, Memorial University of Newfoundland, St John's, Newfoundland and Labrador, Canada

*Correspondence: Benito Pereyra-Alférez, benito.pereyraal@uanl.edu.mx

ORCID: C.A.F.-C., 0009-0009-7552-6537; E.L.-L., 0009-0001-2370-6327; A.S.-R., 0009-0004-5030-0180; Y.H., 0009-0006-2319-0523; A.S.L., 0000-0002-4510-7683, J.L.G.-G. 0009-0007-3299-2664, S.A. M-M 0009-0004-9045-2146, M.E-S. 0000-0002-4179-0251, L.J.G.-W. 0000-0001-7396-9531, and B.P.-A. 0000-0002-3726-0389

Abstract

Background: *Bacillus thuringiensis* is distinguished from *Bacillus cereus* by a plasmid-encoded phenotype, the synthesis of insecticidal and cytocidal proteins, known as Cry and Cyt, respectively, whereas genome-based classification uses chromosomal genes. We asked whether species identity predicts the presence of pesticidal genes in the *B. cereus* group.

Results: Three soil strains producing a parasporal inclusion originally identified as *B. thuringiensis* were compared with seventeen reference genomes. Average nucleotide identity (ANI), digital DNA–DNA hybridization (dDDH) and a core-genome phylogeny resolved all three assignments concordantly: GM18 was *B. toyonensis* (98.87% ANI, 90.5% dDDH, sister to BCT-7112T at 100% bootstrap), IB84 *B. albus* (97.82%; 80.2%), and L4J2 fell in a near-clonal clade with two subsp. *israelensis* references (99.97% ANI). Gene content differed. L4J2, deposited as *B. cereus* group, carried the complete canonical mosquitocidal cassette — Cry4, Cry10, Cry11, Cyt1 and Cyt2 — at named-gene identity and entirely on plasmids; GM18 carried no *cry* or *cyt* gene, sharing almost no toxin content with its own type strain; IB84 carried no classical insecticidal locus, nor did two reference strains deposited as *B. thuringiensis*. Below the species boundary, the toxin families shared by two genomes were unrelated to how closely related they were. Toxin loci were enriched 4.54-fold on plasmids and lay closer to insertion sequences than expected, whereas virulence factors were not. Seven of ten reference strains deposited as *B. thuringiensis*, including the *kurstaki* strains used commercially, are assigned to *B. cereus* genomically.

Conclusions: Published bioassays of the three strains agree with these predictions, although no claim of insecticidal activity is made here. Species identification cannot substitute for direct gene-content profiling in biocontrol prospecting.

Keywords

Bacillus cereus group; *Bacillus thuringiensis*; average nucleotide identity; biological control; comparative genomics; insertion sequences; mobile genetic elements; pesticidal proteins; plasmids.

Background

The *Bacillus cereus* group comprises closely related spore-forming bacteria whose members span a remarkable range of lifestyles, from the mammalian pathogen *Bacillus anthracis* to the invertebrate-specific *Bacillus thuringiensis* and the food-associated *B. cereus sensu stricto* [1, 2]. Their genome nucleotide sequences are close enough that whole-genome comparisons have repeatedly failed to separate the nominal species. As soon as multilocus and then genome-scale data became available, it was argued that *B. anthracis*, *B. cereus* and *B. thuringiensis* constitute a single genomic species on genetic evidence [3], and subsequent genome-based surveys have confirmed that the group behaves as a near-continuum in which the conventional 95–96% average nucleotide identity (ANI) boundary does not align with the names in use [4, 5]. The nomenclatural response has been to propose genome-anchored species within the group [6] and taxonomic schemes that reconcile genomic definitions with the clinical and industrial phenotypes that are relevant to practitioners [5].

The reason for the mismatch is not statistical but mechanistic, and it has been known since the earliest molecular work on the group. What classically defines *B. thuringiensis* is the production of parasporal crystals of insecticidal Cry and Cyt proteins, and the genes encoding those proteins were shown to reside on large plasmids, transmissible between *B. thuringiensis* and *B. cereus* strains by conjugation, four decades ago [7]. The complete *B. thuringiensis* subsp. *israelensis* toxin-coding plasmid pBtoxis was later sequenced and shown to carry the entire dipteran-active gene set [8]. A classification built on the chromosome therefore cannot, even in principle, track a trait carried on the accessory replicons — and the current structure-based nomenclature for these proteins was introduced precisely because the older activity-anchored naming conflated sequence relationships with biological effect [9, 10].

This has practical consequences, because *B. thuringiensis* is the basis of the most widely deployed microbial insecticides in use. Formulations built on subsp. *kurstaki* strains have controlled lepidopteran pests for decades, and subsp. *israelensis* preparations are the mainstay of larval mosquito and blackfly control worldwide, with a dipteran specificity that traces to the combined action of Cry4, Cry10, Cry11, Cyt1 and Cyt2 [11, 12, 13, 14]. The same organisms also yield parasporins, parasporal proteins without insecticidal activity that are selectively cytotoxic to human cancer cell lines [15, 16, 17]. Bioprospecting for new strains in either application commonly proceeds by isolating candidates, identifying them to species, and prioritizing those identified as *B. thuringiensis*. If species identity and pesticidal gene content are inherited through different routes, that workflow has a failure mode in both directions: it will

discard strains carrying the arsenal under another name, and it will promote strains carrying the name without the arsenal.

Whether it does so is an empirical question about how strongly relatedness predicts gene content, and it has not been quantified on a panel assembled to test it. We addressed it with three strains from a soil collection, sequenced and deposited previously, together with seventeen reference genomes chosen to provide species type strains, the subspecies references behind commercial formulations, strains characterized for parasporin production, and *B. thuringiensis* genomes spanning a range of *cry* content. The three strains turned out to cover the three possible combinations of name and arsenal, which allows the question to be addressed both as a panel-wide statistical statement and as three concrete cases. We report their species assignments, the pesticidal gene content of every genome with its sequence-evidence tier and reading-frame integrity declared, the relationship between genomic relatedness and repertoire overlap across all 190 pairs, and the mobilome features that account for what we observe.

Methods

Bacterial strains, isolation and prior characterization

The three strains examined here — GM18, IB84 and L4J2 — belong to the *Bacillus thuringiensis* collection of the Instituto de Biotecnología, Facultad de Ciencias Biológicas, Universidad Autónoma de Nuevo León. All three were recovered from agricultural soil by the heat-treatment and selective-media procedure of Travers *et al.* [18]: GM18 from the state of Guanajuato, IB84 from San Luis Potosí and L4J2 from northern Mexico.

All three were originally identified as *B. thuringiensis* on phenotypic grounds — Gram-positive, sporulating, and producing a parasporal crystal. GM18 was characterized in more detail: based on its H antigen, it was recognized as a new serotype (H-24) and described as *B. thuringiensis* subsp. *neoleonensis*, a subspecies producing a triangular crystal [19]. We follow the spelling of that original description; the genome announcement for this strain renders the epithet "*neolonensis*" [20]. Preliminary bioassays reported in the three genome announcements [21, 20, 22] are used here as comparison data for the gene content we describe and are tabulated in Additional file 2: Table S9. They are qualitative results: no dose–response data, LC50 values or expression measurements exist for any of the three strains, and none were generated here.

Genome sequences analyzed

The genome of the three strains was sequenced and deposited, each with its own announcement [21, 20, 22]. This study analyzes deposited assemblies rather than re-assembling the reads: GCA_015620955.1 for GM18, GCA_015620945.1 for IB84 and GCA_042535225.1 for L4J2. Library preparation, sequencing, and assembly are described in the announcements and are not repeated here. Analyzing the deposited assemblies guarantees that every accession cited is the exact sequence on which the reported values were computed and keeps the analysis downstream of the NCBI contamination screen applied at submission. All assembly statistics in Table 1 were recomputed from the deposited FASTA files rather than taken from the announcements, because for L4J2 the two disagree (see Discussion).

The 17 publicly available comparison genomes are listed in Additional file 2: Table S7 with their assembly accessions and assembly levels. They were retrieved from NCBI either with the datasets command-line tool, for entries carrying a GCA assembly accession, or through the NCBI E-utilities efetch endpoint for entries given only as sets of nucleotide accessions, in which case all replicons of a strain were concatenated into a single per-strain FASTA file. The panel was assembled to place the three study strains against four kinds of anchor: species type strains (*B. cereus* ATCC 14579T [3], *B. toyonensis* BCT-7112T, *B. albus* N35-10-2T, *B. mycoides* DSM 2048T and *B. pseudomycoides* DSM 12442T), reference strains of the *B. thuringiensis* subspecies used in commercial formulations (*kurstaki* HD-1 and HD73, *israelensis* HD-789 and AM65-52, *aizawai* HD133), strains characterized for parasporin production (A1100, ST7, BD59S) and additional *B. thuringiensis* genomes spanning a range of *cry* gene content (BMB171, Bt407, HS18-1, and the *B. wiedmannii* biovar *thuringiensis* strain FCC41 [23]). *B. pseudomycoides* DSM 12442T served as the outgroup throughout.

A note on nomenclature

Throughout this work, *B. thuringiensis* and *B. cereus* were used in their classical phenotype-anchored sense when referring to deposited strain designations; published literature and genome-based species assignments are always attributed explicitly to GTDB R232 [24, 25]. The two systems disagree for a substantial part of the panel, and that disagreement is itself one of our results; we therefore never use a species name without stating which system produced it. GTDB placeholder suffixes of the form _S, which denote a genome-supported species distinct from classical usage, are retained in the supplementary tables and omitted from the figures for legibility.

Assembly quality control

Assembly metrics were recomputed for all 20 genomes using QUAST v5.3.0 [26] with a minimum contig length of 500 bp. Completeness and contamination were estimated with CheckM2 v1.1.0 [27] and, independently, with BUSCO v6.1.0 [28] against the bacteria_odb12.2 lineage dataset (n=116) applied at the domain level; the CheckM estimates produced internally by the annotation pipeline provide a third, independent value. Each deposited assembly was validated by mapping that strain's own public sequencing reads, retrieved from the Sequence Read Archive runs listed in Table 1, back onto it with BWA-MEM2 v2.3 [29], with alignment statistics computed using samtools v1.24 [30]. Read quality was assessed with FastQC v0.12.1 and aggregated with MultiQC v1.35 [31].

Genome annotation

All 20 genomes were annotated uniformly with Bakta v1.10.1 [32] against the full database (schema v5, minor version v5.1), using -genus *Bacillus* and 12 threads. --species and --strain were supplied for the 17 reference genomes, whose taxonomy is established, and deliberately omitted for the three study strains, whose species-level identity is itself a question addressed here. All 20 genomes were additionally annotated using the NCBI Prokaryotic Genome Annotation Pipeline (PGAP, build 2026-06-18.build 8602) [33], which provided a second, independent set of pseudogene calls. Bakta annotations were used for all coordinate-based analyses; PGAP annotations were used only as an independent cross-check.

Taxonomic assignment and species delineation

Genome-based taxonomic assignment used GTDB-Tk v2.7.2 [34] against release R232 [24]; all 20 genomes were classified by ANI screening with skani [35] alone, without requiring phylogenetic placement.

Average nucleotide identity was computed for all pairs applying fastANI v1.34 [36], orchestrated through ANIclustermap v2.0.1 in -mode fastani, giving 190 distinct pairwise values. Species blocks were delimited by single-linkage clustering of the ANI matrix; because single linkage chains genomes through intermediates, it is deliberately permissive and was used to test whether a given threshold separates the panel, not as a formal classification proposal. To assess how far the outcome depends on that choice, the clustering was recomputed independently at each of 96 thresholds spanning 90.0–99.5% ANI.

Digital DNA–DNA hybridization (dDDH) was estimated for all 190 pairs with the Genome-to-Genome Distance Calculator of the DSMZ [37, 38], submitted in three waves. Formula 2, which normalizes identities over the length of the high-scoring segment pairs, was used throughout because it is the way recommended when draft genomes are involved. Species-level interpretation used the conventional thresholds of 95–96% ANI [39, 36, 40] and 70% dDDH [41], following the proposed minimal standards for the use of genome data in prokaryotic taxonomy [42].

Verification of the dDDH result set

Because the GGDC web service labels results by submitted file name, each genome was submitted under a unique name, and the returned table was checked programmatically before use. Two tests were applied. First, an assertion that no two distinct genomes compared against the same reference return an identical triplet of formula 1, 2 and 3 values to the first decimal place, a coincidence that is not possible for distinct genomes; it passed with zero violations across all 190 pairs. Second, a consistency check against ANI, exploiting the tightness of the ANI–dDDH relationship in this panel: a degree-two polynomial fitted to dDDH against ANI leaves a residual standard deviation of 1.112 percentage points, and none of the 54 pairs involving a study strain deviates from that fit by more than three standard deviations. The two criteria are therefore mutually corroborating rather than merely co-reported, and the concordance is shown in Additional file 1: Fig. S1 and tabulated in Additional file 2: Table S1. ANI values are independent of this submission route, having been computed locally.

Phylogenomic inference

A core-genome alignment was built from the uniform Bakta annotations of all 20 genomes with Panaroo v1.8.0 [43] in strict mode with a core threshold of 0.95. A maximum-likelihood phylogeny was inferred with IQ-TREE2 v2.3.6 [44] under the GTR+F+I+R9 substitution model selected by ModelFinder [45] on that alignment, with branch support from 1000 ultrafast bootstrap replicates [46]. The tree was rooted on *B. pseudomycoides* DSM 12442T.

For comparison, a second phylogeny was inferred from a reduced marker set using VBCG v1.3, which selects 20 vertically inherited single-copy core genes and runs an internal workflow of Prodigal [47] for gene calling, HMMER [48] for marker detection, MUSCLE [49] for per-gene

alignment, Gblocks [50] for trimming, concatenation, and RAxML [51] for inference with 100 bootstrap replicates. No outgroup was specified at runtime, so the distributed tree is arbitrarily rooted; we rooted it on the same outgroup to permit comparison. Topological distance between the two phylogenies was quantified as the Robinson–Foulds distance [52] using dist.topo(method="PH85") in ape v5 [53], computed on the unrooted trees so that our own rooting choice could not inflate the measure.

Pangenome analysis

The pangenome was built with Panaroo v1.8.0 [43] in strict mode with a core threshold of 0.95 and partitioned independently with BPGA v1.3 [54] using USEARCH clustering at a 50% identity cutoff. Strain-specific genes were defined as Panaroo gene clusters present in exactly one genome of the panel; BPGA's independent definition of unique genes was used as a contrast. Functional annotation of the 19 909 pangenome gene-cluster representatives used eggNOG-mapper v2.1.15 [55] in diamond mode against eggNOG v5.0.2 [56], which annotated 16 257 clusters (81.7%). Orthologous group category letters were collapsed into the four conventional functional divisions; a gene carrying several letters contributes to each, so the denominator for category proportions is the number of assignments rather than the number of genes.

Detection and confidence tiering of pesticidal protein genes

Pesticidal protein genes were detected with BtToxin_Digger v1.0.10 [57] in nucleotide mode, which performs six-frame translation of each assembly, retains open reading frames of at least 75 codons, and combines homology search, hidden Markov model profiles and a support-vector-machine classifier. Every locus was verified independently by BLASTP and TBLASTN (BLAST+ v2.16.0 [58]; HMMER v3.4 [48]) against a local copy of the Bacterial Pesticidal Protein Resource Center database (downloaded on 23 September 2026) [10], and the two sets of calls were reconciled locus by locus. All 20 genomes completed this pipeline under identical settings, with no genome excluded or handled separately, so that the counts reported for the three study strains and for the 17 references are directly comparable. Loci were named under both the classical Cry/Cyt/Vip nomenclature [9] and the current structure-based App/Bmp/Gpp/Mpp/Mtx/Sip/Spp/Tpp/Xpp nomenclature [10]; because the two schemes classify partly overlapping sets of proteins, 16 loci are counted under both and the two class totals must not simply be added (see Results). Loci were aggregated to family level with the regular expression ^([A-Za-z]+[0-9]*), so that for example Cry4Ca contributes to family Cry4 and Vip3Aa to family Vip3.

Every locus was assigned to one of four tiers of sequence evidence by the identity of its best database match: named gene, ≥95% identity; same subclass, 78–95%; same family, 45–78%; and distant match, <45%. A fifth category records loci detected only by hidden Markov model profile, with no homology match. These tiers are thresholds on sequence identity against a reference database. They are not measures of biological activity, and no claim of insecticidal activity is made anywhere in this work.

Assessment of coding integrity

All 271 detected locus entries were assessed for reading-frame integrity by direct translation: locus length was checked for being a multiple of three, internal stop codons were sought, and the length of the resulting protein was compared with that of its reference. Strand was resolved independently per locus, because the reconciled call tables retain only minimum and maximum coordinates. The translation result was cross-checked against the two independent pseudogene call sets from Bakta [32] and PGAP [33]. Each locus received a single verdict, evaluated in the following order: internal stop codon detected; pseudogene confirmed by PGAP; pseudogene flagged by Bakta; fragment at a contig edge; otherwise, intact. The contig-edge category is an assembly artefact rather than a biological observation and is reported separately for that reason.

Replicon assignment and mobilome analysis

Every locus was assigned to a chromosome or a plasmid. For the 14 closed genomes of the panel the assignment was read from the replicon declarations in the sequence records themselves; for the six draft genomes — the three study strains together with HD-1, HD133 and N35-10-2T — contigs were classified with RFPlasmid [59] using its *Bacillus* model. MOB-suite v3.1.9 [60] provided a second, independent assignment for cross-checking, and Bakta supplied locus coordinates and flanking gene context. Every locus entry received an assignment; the distinction that matters is therefore not determined versus undetermined but declared (in complete genomes, where the replicon is stated in the sequence record) versus predicted (in draft genomes), and the figures distinguish the two.

Plasmids were reconstructed, typed and assigned a mobility class with MOB-suite v3.1.9 [60], which classifies a plasmid as conjugative, mobilisable or non-mobilisable according to the presence of a relaxase, a type IV secretion system and an origin of transfer. Independent replicon typing used PlasmidFinder [61] run through ABRicate, and plasmid families were clustered with skani [35]. MOB-suite failed to recover 4 of the 40 known plasmids of at least 10 kb present in the complete reference genomes, so all plasmid counts are reported as lower bounds.

Prophages were identified independently with geNomad v1.12.0 [62] and PhiSpy v5.0.10 [63] using its *Bacillus* training model, and the completeness of each predicted region was assessed with CheckV v1.1.1 [64]; regions of High or Complete quality supported by both callers were retained as integrated prophages. Insertion sequences were detected with ISEScan v1.7.3 [65], named by BLASTN against ISfinder [66], and complemented by a BLASTN search against TnCentral [67] for transposons outside the insertion-sequence class. Synteny of the two toxin loci shared across genomes was compared with clinker [68].

Defense systems and CRISPR spacer targeting

Antiviral defense systems were identified with DefenseFinder v3.0.0 [69] and PADLOC v2.0.0 [70], and only systems surviving reconciliation of both outputs were retained. CRISPR-Cas loci were typed with CRISPRCasTyper v1.9.0 [71] and repeats detected independently with CRISPRDetect v2.4 [72]. Spacers extracted by both tools were searched against the plasmids and prophages carried by the other genomes of the panel using CRISPRTarget [73] and BLASTN [58]; only matches supported by both methods and reaching statistical significance were retained, matches against a genome’s own elements were excluded, and a shuffled-spacer control

was run to bound the false-positive rate. Spacer matches document that a lineage has encountered an element resembling one carried by another genome today; they establish neither the direction nor the date of any transfer [74].

Biosafety-relevant gene content

Virulence factors were detected with ABRicate v1.4.0 against VFDB [75] (database version of 3 April 2026), and matches were grouped into functional clusters after filtering on identity and coverage. Antimicrobial resistance genes were detected with ABRicate v1.4.0 against all five bundled databases — CARD [76], ResFinder [77], ARG-ANNOT [78], MEGARes [79] and the NCBI reference gene catalogue [80] — with calls reconciled by genomic coordinate so that a locus detected by several databases is counted once, and the number of supporting databases recorded per locus. Biosynthetic gene clusters were identified with antiSMASH v8.0.4 [81] using the full database, and bacteriocins by BLASTP against the 821-sequence class I, II and III reference protein panel of BAGEL4 [82]; the BAGEL4 orchestration tool itself was not used. Presence of the cereulide synthetase cluster [83] was checked explicitly as a biosafety control. Detection of a resistance or virulence gene is not evidence of the corresponding phenotype, which can be established only by susceptibility testing or infection assay.

Statistical analysis

Two randomization tests were used. Plasmid enrichment of each category of genomic element was assessed by a Monte Carlo test in which the loci of that category were reshuffled at random across the replicons of their genome in proportion to replicon length, and the observed proportion on plasmids was compared with the resulting null distribution. The test is one-sided and designed to detect enrichment; a point estimate below unity therefore licenses the description “not enriched” and not “significantly depleted”.

Proximity of insertion sequences to each class of locus was assessed by comparing the observed number of loci with at least one insertion sequence within a given window against the expectation under random placement within the same replicon, which controls for the differing baseline element density of chromosomes and plasmids. Windows of 1, 5 and 10 kb were examined; the three are nested and therefore not independent. This analysis was restricted to the 14 complete genomes, because contig edges prevent reliable distance measurement in drafts, which means the three study strains do not contribute to it directly.

Repertoire overlap between genomes was quantified as the Jaccard index over sets of toxin families, computed on family presence and absence without weighting by locus count or evidence tier. The index is unstable for small repertoires, and pairs in which either genome carries fewer than five families are marked separately in the figures and are not interpreted as biological similarity. Trends against ANI were summarized with local regression (loess) and a 95% confidence band, which is descriptive and must not be read as prediction. All statistical analysis and all figures were produced in R v4.3.3 with ggplot2, ape v5 [53], dplyr, tidyr, patchwork, ggrepel, scales and svglite.

Use of large language models

Large language model assistants were used during preparation of this manuscript for two purposes: language editing of the English text, and consistency auditing — systematically cross-

checking that numerical values, locus counts, accession numbers and stated claims were reported identically in the text, tables and figure legends. No analysis was performed by a language model. All genomic analyses were executed with the software listed in Additional file 2: Table S8, and no sequence, result or statistic reported here was generated, imputed or altered by a language model. The authors verified every reported value against the primary output files and take full responsibility for the content of this work.

Results

Genome sequences of the three study strains

The three genomes spanned 5.97–7.02 Mb with a G+C content of 34.9–35.1 mol% and 5959–7063 coding sequences called by Bakta, values that sat inside the range spanned by the 17 reference genomes of the panel (Table 1). CheckM2 estimated all three complete [27], and BUSCO recovered 100% of the 116 bacterial single-copy orthologs in IB84 and L4J2 and 99.1% in GM18 [28]. Between 99.65% and 99.90% of each strain's own sequencing reads are mapped back to its assembly.

Table 1. Genome characteristics and species assignment of the three strains sequenced in this study. Assembly statistics from QUAST v5.3.0 (contigs ≥500 bp) and Bakta v1.10.1 (all contigs); completeness from CheckM2 v1.1.0 and BUSCO v6.1.0 against bacteria_odb12.2 (n=116). ANI from fastANI v1.34, genomic classification from GTDB-Tk R232, and core-genome sister taxon and support from the phylogeny of Fig. 2. GM18's dDDH value against its type strain is among those affected by the submission labelling failure described in Methods and is not used to support its assignment.

Characteristic	GM18	IB84	L4J2
Isolation source	agricultural soil, Guanajuato, Mexico	agricultural soil, San Luis Potosi, Mexico	agricultural soil, northern Mexico
Original identification	*B. thuringiensis* subsp. *neoleonensis*, serotype H-24	*B. thuringiensis*	*B. thuringiensis*
Published preliminary bioassay	no insecticidal activity; inclusion proteins cytotoxic to cancer cell lines	crystal proteins not toxic to insects	crystal proteins lethal to Aedes aegypti larvae
Genome announcement	Moreno-Pena et al. 2021	Garcia-Garcia et al. 2021	Rangel-Heredia et al. 2025
Assembly accession analysed	GCA_015620955.1	GCA_015620945.1	GCA_042535225.1

Characteristic	GM18	IB84	L4J2
BioProject / SRA run	PRJNA676032 / SRR13402414	PRJNA676033 / SRR13427799	PRJNA1130484 / SRR29744432
Genome size (bp)	6,129,014	5,965,175	7,016,826
Contigs, all / >=500 bp	156 / 114	156 / 117	799 / 211
Largest contig (bp)	588,862	531,084	856,121
N50 / L50 (>=500 bp)	171,900 / 10	155,983 / 12	349,725 / 6
G+C content (mol%)	34.9	35.1	35.0
CDSs (Bakta / PGAP)	6,074 / 6,244	5,959 / 6,112	7,063 / 7,657
tRNAs / rRNAs (Bakta)	96 / 12	80 / 6	122 / 19
Pseudogenes (Bakta / PGAP, % of CDS)	44 / 300 (4.95)	42 / 243 (3.92)	26 / 297 (3.79)
CheckM2 completeness / contamination (%)	100.0 / 0.85	100.0 / 0.05	100.0 / 5.97
PGAP CheckM completeness / contamination (%)	99.41 / 0.22	99.38 / 0.00	98.82 / 9.07
BUSCO complete (%)	99.1	100.0	100.0
Reads mapped to own assembly (%)	99.65	99.80	99.90
Closest relative by ANI	BCT-7112T (*B. toyonensis* type strain)	N35-10-2T (*B. albus* type strain)	HD-789 (Bt subsp. israelensis)
ANI to closest relative (%)	98.87	97.82	99.97
Margin to second-closest genome (pp)	7.29	4.11	0.01 (AM65-52, same clade)
dDDH to type strain (%)	90.5 (BCT-7112T)	80.2 (N35-10-2T)	65.4 (ATCC 14579T)

Characteristic	GM18	IB84	L4J2
GTDB R232 assignment (skani ANI %)	*B. toyonensis* (99.07)	*B. albus* (98.12)	*B. thuringiensis*_S (99.11)
Core-genome sister taxon (UFBoot)	BCT-7112T (100%)	N35-10-2T (100%)	AM65-52 (88%)
Marker-gene tree sister taxon (bootstrap)	BCT-7112T (100%)	N35-10-2T	HD-789 (99%)
Toxin families / distinct loci	10 / 18	3 / 3	12 / 24
Classical insecticidal loci	3 (all Vpb1, <45% identity)	0	14
Structure-based (parasporin) loci	15	3	10
Loci with intact reading frame	12/18 (67%)	3/3	11/24 (46%)
Plasmid fraction of genome (%)	20.7	22.7	26.3
Conjugative plasmids (>=10 kb)	3	3	2
IS elements (per Mb)	96 (13.71)	54 (8.38)	63 (7.41)
IS per Mb, chromosome vs plasmid	2.26 vs 57.47	1.97 vs 30.49	2.13 vs 22.25
Integrated prophages (% of genome)	2 (0.87)	1 (0.31)	10 (3.98)
Defence systems (on plasmid)	25 (1)	16 (0)	18 (0)
Complete CRISPR-Cas (subtype)	yes (I-C)	no	no
AMR loci (on plasmid)	7 (1)	9 (2)	11 (4)
BGC regions	13	13	17
Virulence-factor clusters	5	8	8

Assembly quality was not uniform across the three, and the differences constrained different parts of the analysis. GM18 and IB84 were comparably contiguous — 114 and 117 contigs of at least 500 bp, with N50 values of 172 kb and 156 kb — and both were essentially free of contamination (CheckM2 0.85% and 0.05%). L4J2 was the exception on both counts. It was the most fragmented assembly of the whole panel, with 799 contigs, and its contamination estimate exceeded the 5% ceiling that defines a high-quality draft [84]: CheckM2 returned 5.97%, and the independent CheckM estimate produced by the PGAP pipeline returned 9.07%, the highest value in the panel. L4J2 nonetheless had the highest N50 of the three (350 kb) and the largest single contig (856 kb), so the fragmentation was concentrated in short contigs rather than distributed across the assembly.

These two observations bound the interpretation of L4J2's size. L4J2 carried approximately 0.9 Mb more sequence than GM18 and 1.1 Mb more than IB84, and its G+C content matched theirs, which argues against gross foreign contamination. But with two independent contamination estimates above the high-quality threshold and 799 contigs, part of that excess may be redundant or extraneous sequence rather than additional genuine content. We therefore treat L4J2's genome size, and every count that scales with it, as an upper bound pending a closed assembly, and we do not use its size as evidence for anything.

All three strains receive unambiguous species assignments

Across the 190 pairwise comparisons of the panel, ANI ranged from 83.09% to 99.98% with a median of 91.58% (Fig. 1). Single-linkage clustering at the conventional 95% threshold recovered seven blocks, and *B. pseudomycoides* DSM 12442T reached no more than 83.7% ANI against any other genome, confirming quantitatively that it is a suitable outgroup.

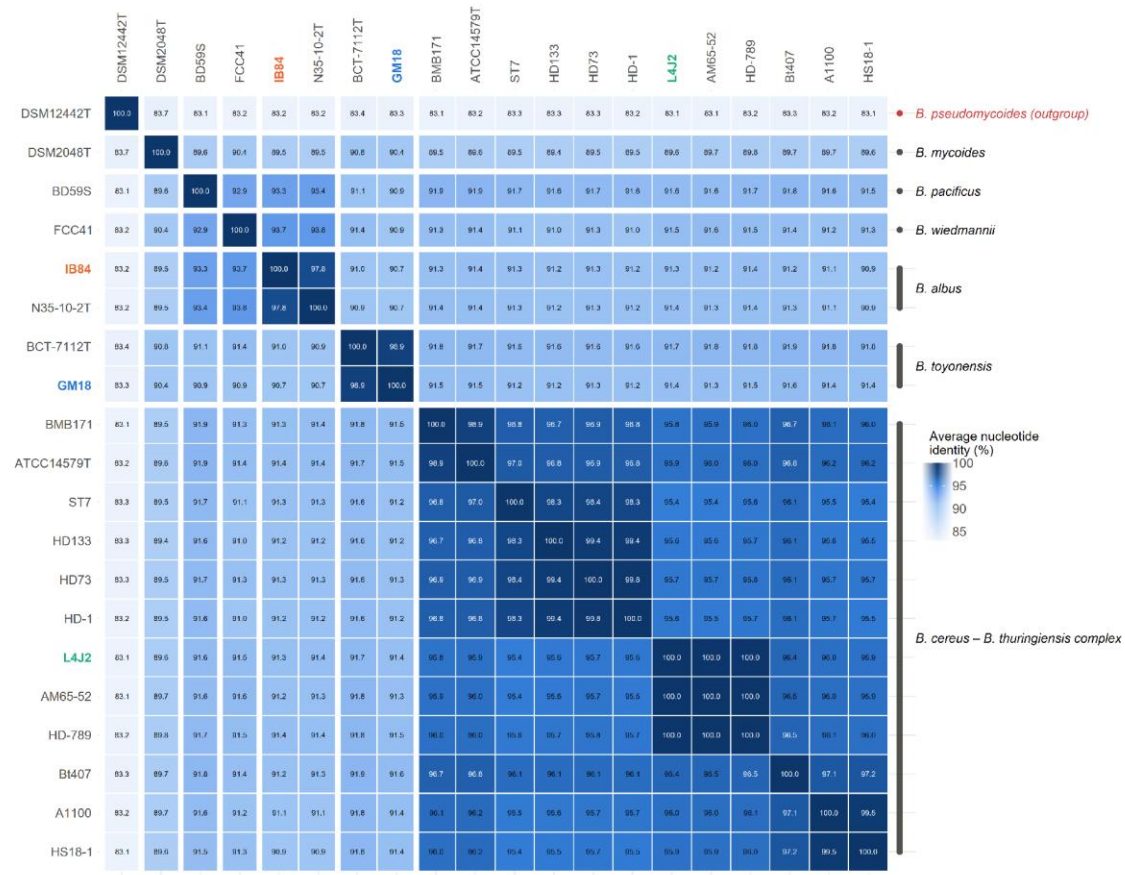


Fig. 1. Pairwise average nucleotide identity among twenty *Bacillus cereus* group genomes. Each cell gives the percentage ANI between the genomes of that row and column, computed with fastANI v1.34. Rows and columns are ordered by UPGMA hierarchical clustering on the distance 100 − ANI, so that related genomes are adjacent. The colour ramp is sequential from 83% to 100% with its inflection fixed at 95%, the conventional prokaryotic species threshold. Brackets on the right delimit the blocks obtained by single-linkage clustering at that threshold and are labelled with the corresponding species name; the red bracket marks the outgroup *Bacillus pseudomycoides* DSM 12442T. Coloured strain labels indicate the three strains sequenced in this study. The three values displayed as 100.0 in the L4J2 / AM65-52 / HD-789 trio are 99.97% rounded to one decimal place; these are not identical genomes.

Each study strain had a single, well-separated closest relative (Fig. 1, Additional file 1: Fig. S2). GM18 shared 98.87% ANI with BCT-7112T, the type strain of *B. toyonensis* [85], and its second-closest genome, Bt407, lay 7.29 percentage points further away at 91.58%. IB84 shared 97.82% ANI with N35-10-2T, the type strain of *B. albus* [6], and its second neighbor, FCC41, lay 4.11 points away at 93.71%. L4J2 was a different case: it shared 99.97% ANI with HD-789

and effectively the same value with AM65-52 (a margin of 0.01 points), two reference strains of *B. thuringiensis* subsp. *israelensis*, so the three were near-clonal, and no single nearest relative can be identified among them. None of the three strains fell in the 95–96% band in which the ANI criterion is least reliable [39, 36, 42]. The margin to the second neighbor matters as much as the absolute value: a high ANI with a narrow margin would indicate a continuum of divergence rather than a discrete species. For GM18 and IB84, that margin was wide, and the ANI step at the block boundary was abrupt. For L4J2, the margin was negligible because its two closest relatives were themselves near-identical to it, which is why we describe L4J2 by its clade rather than by a nearest-neighbor species name. GTDB-Tk R232 returned the same three assignments independently: *B. toyonensis*, *B. albus*, and *B. thuringiensis* within the *B. cereus* group, at closest-genome ANI values of 99.07%, 98.12%, and 99.11% respectively [34, 24].

Digital DNA–DNA hybridization resolved all 190 pairs of the panel and agrees with the ANI assignments throughout. GM18 against the *B. toyonensis* type strain gave 90.5%, and IB84 against the *B. albus* type strain gave 80.2%, both far above the 70% species threshold [41]. L4J2 against the *B. cereus* type strain ATCC 14579T gave 65.4%, below that threshold, which is consistent with its GTDB assignment to a *thuringiensis* clade rather than to *B. cereus sensu stricto* and is a further reason to describe it by clade. The two criteria were quantitatively consistent across the panel: a degree-two polynomial fitted to dDDH against ANI left a residual standard deviation of 1.112 percentage points, and none of the 54 pairs involving a study strain fell outside three standard deviations of that fit (Additional file 1: Fig. S1, Additional file 2: Table S1).

The core-genome phylogeny, inferred from 3399 filtered core gene alignments under GTR+F+I+R9, was essentially fully resolved: 16 of its 17 internal branches reached 100% ultrafast bootstrap support (Fig. 2) [43, 44, 45, 46]. GM18 was the sister taxon of BCT-7112T and IB84 the sister taxon of N35-10-2T, both at 100% support. L4J2 fell inside a near-clonal clade with AM65-52 and HD-789, with terminal branch lengths of the order of 10-6 substitutions per site. The single branch below 100% was the bifurcation separating L4J2 from AM65-52, at 88%, and it should not be read as uncertainty about where L4J2 sits: its membership of the clade was supported at 100%, and the 88% reflected the absence of phylogenetic signal to order two genomes that share 99.97% ANI. The 95% ANI blocks corresponded to monophyletic clades of this tree without exception.

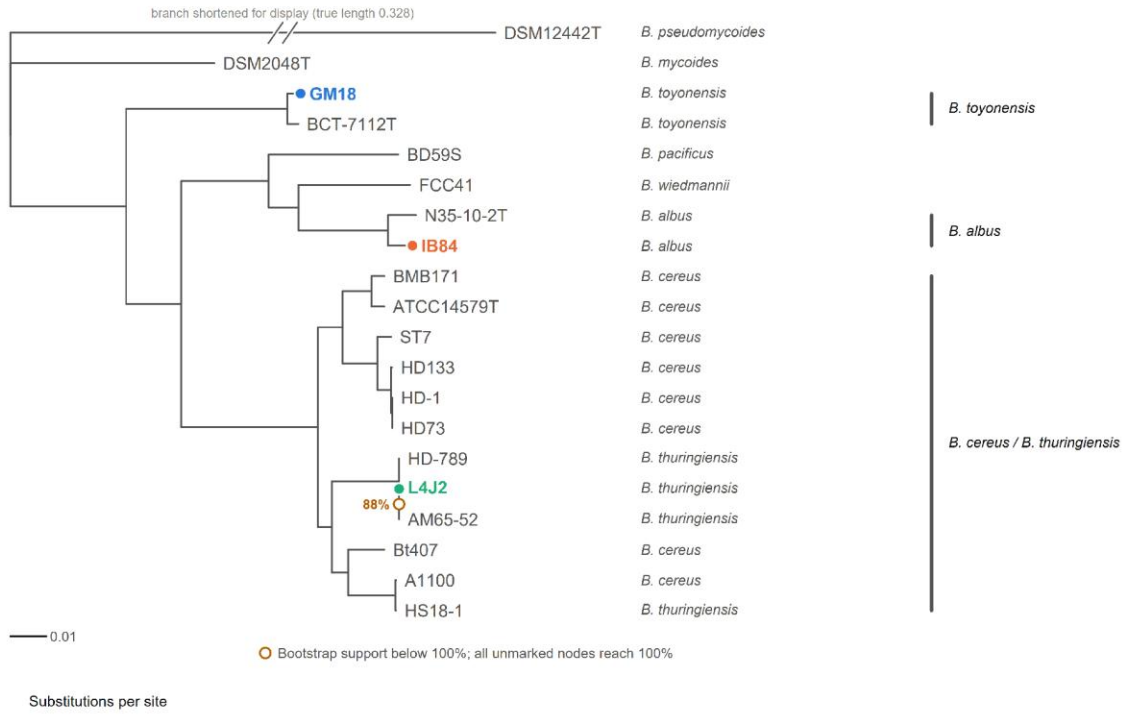


Fig. 2. Core-genome maximum-likelihood phylogeny of twenty *Bacillus cereus* group genomes. Core-genome alignment built with Panaroo v1.8.0 and inference with IQ-TREE2 v2.3.6 under the GTR+F+I+R9 model selected by ModelFinder, with 1000 ultrafast bootstrap replicates; the tree is rooted on *Bacillus pseudomycoides* DSM 12442T. Support values are shown only at nodes below 100%: the sole exception is the split separating L4J2 from AM65-52 (88%), which distinguishes two genomes sharing 99.97% ANI and does not indicate uncertainty about L4J2's placement, since its membership of the clade is supported at 100%. Coloured tips mark the three strains sequenced in this study; italics give the species assigned by genome-based classification and the brackets on the right the blocks delimited by ANI at 95%. The outgroup branch is graphically shortened to preserve legibility and its true length is printed at the break mark; any distance measured on the figure in that region is not the real one. Scale bar, 0.01 substitutions per site.

A second phylogeny inferred from 20 vertically inherited marker genes placed all three study strains correctly — GM18 as the direct sister of BCT-7112T at 100% bootstrap support, IB84 with N35-10-2T, and L4J2 with HD-789 at 99% — but it did not recover the topology of the core-genome tree as a whole (Additional file 1: Fig. S3, Fig. S4). Four internal nodes were shared between the two trees, giving a Robinson–Foulds distance of 26 out of a possible 34 [52]. Two indications identify this as loss of resolution rather than an alternative evolutionary history: the inference declared three genome pairs *exactly identical* under this marker set (HD73 ≡ HD−1, L4J2 ≡ HD−789, HS18−1 ≡ A1100), and the discrepancies concentrate in precisely the regions of the tree where divergence is lowest. A third, stronger indication emerged from re−analysis. When

one assembly of the panel was replaced by a more contiguous version of the same strain, the core-genome topology did not change at all (Robinson–Foulds distance 0, all 17 splits preserved), whereas the marker-gene topology moved by a Robinson–Foulds distance of 16. A 20-gene sample is therefore sensitive to assembly input in a way that a 3399-gene core alignment is not, and we report the marker tree as a limit of that approach rather than as independent corroboration.

The 95% species boundary does not resolve the *B. cereus* group, and the deposited name does not predict the genomic assignment

Varying the clustering threshold from 90.0% to 99.5% ANI showed that the conventional cut landed in a flat region rather than at a natural break (Additional file 1: Fig. S5). At 95%, a single cluster absorbed 12 of the 20 genomes — A1100, AM65-52, ATCC 14579T, BMB171, Bt407, HD-1, HD-789, HD133, HD73, HS18-1, L4J2 and ST7 — placing the type strain of *B. cereus*, several *B. thuringiensis* reference strains and one study strain in the same putative species. The total number of clusters remained fixed at seven throughout 93.8–96.5%: the panel did not reorganize anywhere near the threshold in general use. GM18 and IB84 were entirely robust to this choice, persisting as clean two-genome clusters with their respective type strains; L4J2 was the opposite case, and its assignment to one nominal species or another depended wholly on the cut applied. We therefore describe L4J2 by its clade and its gene content rather than by a species name that the criterion does not sustain.

The consequence for nomenclature was direct (Fig. 3, Additional file 2: Table S4). Of the ten reference strains deposited as *B. thuringiensis*, seven were assigned to *B. cereus* on genomic criteria: HD-1, HD73, BMB171, Bt407, HD133, A1100 and ST7. Only HD-789, AM65-52 and HS18-1 retained a *thuringiensis* designation. Eight of the 17 reference genomes (47%) showed some disagreement between deposited name and genomic assignment, and BD59S, deposited without a species epithet, was assigned to *B. pacificus*. The conflict was sharpest in the best-characterized strains in the field: HD-1 and HD73 are the subsp. *kurstaki* strains that have underpinned commercial *Bt* formulations for decades [11, 12] and HD133 is the subsp. *aizawai* reference, and all three were called *B. cereus* here. None of the three study strains showed any such conflict.

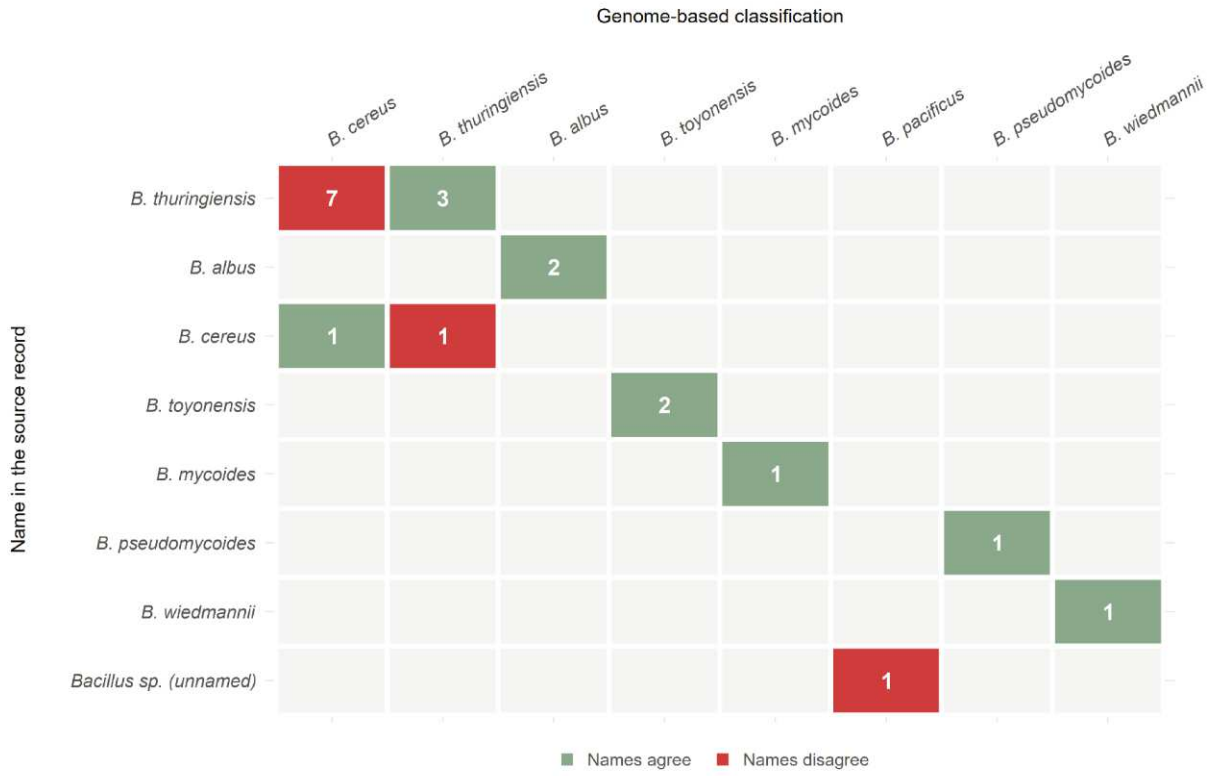


Fig. 3. Deposited species names versus genome-based classification. For each of the twenty genomes, the name under which it is deposited or described in the literature (left) is set against the name assigned by GTDB-Tk R232 (right). Red lines mark genomes where the two disagree at species level: seven of the ten reference strains deposited as *B. thuringiensis*, among them the *kurstaki* strains HD-1 and HD73 used in commercial formulations, are assigned to *B. cereus* on genomic criteria. The three strains sequenced in this study receive consistent names in both systems. Neither column is the correct one: the left encodes a phenotype-anchored nomenclature with historical priority, the right an operational classification of the chromosome. The comparison is made at species level; the subspecies and serovars of the left column have no equivalent in the genomic system, which recognises no infraspecific ranks. GTDB placeholder suffixes of the form _S are omitted here and retained in Additional file 2: Table S4.

Neither of the two classifications compared in Fig. 3 is the correct one. The deposited names encode a phenotype-anchored nomenclature with historical priority; the genomic assignment is an operational classification of the chromosome. That the two disagree for *B. cereus* and *B. thuringiensis* has been argued since whole-genome comparisons first became available [3, 4, 5], and the reason is mechanistic: the trait that classically separates the two species, production of parasporal insecticidal crystals, is encoded on plasmids [7, 8], which the chromosomal classification does not see. The remainder of this section quantifies that explanation on our own panel.

Gene content analysis of the three focal strains

The three strains cover the three possible combinations of name and arsenal (Fig. 4, Additional file 1: Fig. S6, Fig. S7, Table 2, Additional file 2: Table S2).

Table 2. Pesticidal protein gene families of the three study strains, with sequence evidence and replicon location. Families are reported separately under the classical Cry/Cyt/Vip nomenclature and the current structure-based nomenclature; because the two schemes classify partly overlapping sets of proteins, locus counts must not be summed across the two panels. Strongest sequence evidence is the highest identity tier reached by any locus of that family in that genome, against the BPPRC database. Replicon location for these three strains is a prediction from contig classification, not a declaration in a sequence record, and is least reliable for L4J2 given its 799-contig assembly.

Strain	Nomenclature panel	Family	Loci	Strongest sequence evidence	Loci on plasmid	% on plasmid
GM18	classical Cry/Cyt/Vip	Vpb1	3	distant (<45%)	2	67
GM18	structure-based	App4	1	named (>=95%)	1	100
GM18	structure-based	Bmp1	2	distant (<45%)	2	100
GM18	structure-based	Mpp45	1	distant (<45%)	1	100
GM18	structure-based	Mpp75	1	distant (<45%)	1	100
GM18	structure-based	Mpp93	1	distant (<45%)	1	100
GM18	structure-based	Spp1	5	same subclass (78-95%)	3	60
GM18	structure-based	Tpp78	1	distant (<45%)	1	100
GM18	structure-based	Vpb4	1	distant (<45%)	1	100
GM18	structure-based	Xpp22	2	distant (<45%)	2	100
IB84	structure-based	Spp1	1	same subclass (78-95%)	1	100

Strain	Nomenclature panel	Family	Loci	Strongest sequence evidence	Loci on plasmid	% on plasmid
IB84	structure-based	Tpp78	1	distant (<45%)	1	100
IB84	structure-based	Xpp22	1	distant (<45%)	1	100
L4J2	classical Cry/Cyt/Vip	Cry10	1	named (>=95%)	1	100
L4J2	classical Cry/Cyt/Vip	Cry11	1	named (>=95%)	1	100
L4J2	classical Cry/Cyt/Vip	Cry4	8	named (>=95%)	8	100
L4J2	classical Cry/Cyt/Vip	Cry69	1	same family (45-78%)	1	100
L4J2	classical Cry/Cyt/Vip	Cyt1	2	named (>=95%)	2	100
L4J2	classical Cry/Cyt/Vip	Cyt2	1	named (>=95%)	1	100
L4J2	structure-based	Mpp60	2	named (>=95%)	2	100
L4J2	structure-based	Mtx1	2	distant (<45%)	2	100
L4J2	structure-based	Spp1	1	same subclass (78-95%)	0	0
L4J2	structure-based	Vpb4	1	distant (<45%)	1	100
L4J2	structure-based	Xpp22	3	distant (<45%)	3	100
L4J2	structure-based	Xpp86	1	distant (<45%)	1	100



Fig. 4. Insecticidal toxin gene families carried by each genome. Presence of the thirty Cry, Cyt, Vip, Vpa and Vpb protein families detected in the panel, identified with BtToxin_Digger v1.0.10 and verified independently by homology search against the BPPRC database. The number in each cell gives the loci of that family present in that genome and the colour the strongest sequence evidence reached, from distant match (<45% identity) to named gene (≥95%); grey marks absence of the family. Rows follow the order of the core-genome phylogeny and coloured labels mark the three strains sequenced in this study. L4J2 carries the Cry4–Cry10–Cry11–Cyt1–Cyt2 cassette characteristic of mosquitocidal activity at named-gene identity, identical in composition to that of AM65-52 and HD-789; IB84 carries no gene of this class at all, and GM18 carries only three Vpb1 loci, all of them distant matches. Colour reflects sequence identity, not insecticidal activity: a named locus does not imply a functional protein, and reading-frame integrity is assessed separately in Fig. S8. Family-level aggregation groups subclasses that may differ in activity spectrum.

L4J2 carried the complete canonical mosquitocidal cassette. It had 14 loci in the classical insecticidal panel. Thirteen of them belonged to five families — Cry4 (eight loci), Cry10 (one), Cry11 (one), Cyt1 (two) and Cyt2 (one), and every one of the 13 reached named-gene identity (≥95%); the fourteenth was a Cry69 locus at family-level identity.

This combination defines dipteran–larvicidal activity in *B. thuringiensis* subsp. *israelensis* [13, 14]. The two *israelensis* reference genomes of the panel, AM65-52 and HD-789, carried the same five families, and at subclass resolution all three strains carried the same seven proteins: Cry4Aa, Cry4Ba, Cry10Aa, Cry11Aa, Cyt1Aa, Cyt1Ca and Cyt2Ba. This was not a diffuse resemblance but an identical repertoire in a genome sharing 99.97% ANI. A screen guided by the deposited species name would have discarded L4J2, which is deposited as *B. cereus* group.

GM18 carried neither *cry* nor *cyt* genes at all. It carried just three *loci* in the classical panel that belonged to the Vpb1 family, a component of a vegetative binary toxin, and all three were distant matches below 45% identity. The distinction matters, and it corrects a framing used earlier in this project: "GM18 carries toxin genes" and "GM18 carries *cry* genes" are not interchangeable statements. In the structure-based panel, GM18 carried 15 *loci* across the widest set of families of the three strains — the second-highest count in the panel after A1100 — including one App4 locus at named-gene identity, which is notable in a *B. toyonensis*, a species absent from the classical parasporin literature. The rest of that load was almost entirely low confidence (Additional file 1: Fig. S7): five Spp1 *loci* at subclass identity and nine distant matches below 45%. The defensible statement is therefore that GM18 carried one identified App4 locus, five Spp1 loci that carry no discriminating information for the reason given below, and nine distant candidates — not that it has a high parasporin load.

IB84 carried no classical insecticidal locus whatsoever, despite having come from sampling directed at *B. thuringiensis*. It was not alone in this: five of the 20 genomes carried no classical insecticidal locus, the others being the *B. cereus* type strain ATCC 14579T, the unnamed isolate BD59S, and BMB171 and Bt407, two strains deposited as *B. thuringiensis* that are themselves acrystalliferous or Cry-negative references and were included in the panel for that reason. IB84 functioned as an internal negative control: a close, non-toxigenic relative within the same group, whose company in that category includes two genomes bearing the *thuringiensis* epithet. Its three structure-based loci comprised the near-ubiquitous Spp1 and two distant candidates.

Two panel-wide observations condition how all of these counts should be read. First, the two classes differed sharply in the strength of the evidence behind them (Additional file 1: Fig. S7): 63 of the 138 entries in the classical insecticidal panel (45.7%) reached named-gene identity, whereas in the 133-entry structure-based panel only a small minority did and the large majority were distant matches below 45% identity. A count of parasporin-like loci is therefore largely a count of candidates, whereas a count of classical insecticidal loci means roughly what it appears to; any figure quoted for the former must state the identity threshold applied. Second, the structure-based family Spp1 was present in 19 of the 20 genomes — the sole exception being BCT-7112T — at a consistent 78–84% identity, including in all three genomes used as negative controls for the classical panel. It carried essentially no discriminating information and was excluded from presence/absence interpretation, being retained only in the raw call table. Across the panel, the highest raw counts were also the least informative: ST7 led with 20 classical loci but concentrated most of them in a single low-confidence family, while HD133's named Cry1 loci are exactly what a subsp. *aizawai* reference should have. Sixteen loci carried a classification under both nomenclatures, so locus totals must not be summed across the two panels: the 271 locus entries of the two panels together corresponded to 243 distinct loci, and the 16 dual-

classified loci do not by themselves account for that difference, so the merge must also collapse duplicate calls within a panel. We report both totals and do not interpret the gap between them.

Toxin-locus integrity differs between the three strains, and for different reasons

Of the 271 locus entries assessed, 170 had an intact reading frame, 57 were complete but truncated relative to their reference, 21 were fragments at a contig edge, 16 were pseudogenes confirmed independently by both annotation pipelines, four were flagged by one annotation only, two had no coding sequence called, and one carried an internal stop codon (Additional file 1: Fig. S8a, Fig. S8b).

The three study strains separate cleanly on this measure. IB84 had all three of its loci intact — on a denominator of three, which makes the percentage statistically fragile, reason why we reported it as 3/3 rather than 100%. GM18 retained 12 of 18 loci (67%) and L4J2 11 of 24 (46%).

For GM18, the integrity deficit was biological rather than an artefact of assembly, and the re-analysis established this directly. The strain was re-examined on a markedly more contiguous assembly of the same genome — 156 contigs instead of 1193, with the N50 rising from 23 kb to 172 kb — and the pseudogene load did not move: Bakta called 44 pseudogenes on both assemblies, and the PGAP pseudogene fraction went from 4.93% to 4.95%. Contiguity was therefore not what limited GM18's intact-ORF fraction, and long-read sequencing would not be expected to recover the missing reading frames. Five of GM18's six incomplete loci were nonetheless contig-edge fragments, so the particular assignment of individual loci to failure modes remains provisional even though the aggregate does not.

For L4J2 the pattern was different in kind. Its losses included pseudogenes confirmed independently by both annotation pipelines alongside six contig-edge fragments, in the genome that carried by far the largest integrated prophage burden of the three (10 integrated prophages occupying 3.98% of the genome) and the most fragmented assembly of the panel. Both the biological and the technical contributions were therefore larger in L4J2 than in the other two strains, and they cannot be separated on these data.

The applied consequence is specific. L4J2's mosquitocidal cassette was present at named-gene identity, but fewer than half of its total toxin loci had an intact reading frame. An intact reading frame is necessary but not sufficient for protein production — it says nothing about expression, secretion or crystal formation.

Published bioassays are concordant with the genomic predictions

Preliminary phenotypic characterizations of all three strains were reported in their genome announcements, and they agree with the gene content described here (Additional file 2: Table S9). L4J2, the only one of the three carrying the mosquitocidal cassette, was the only one reported to kill larvae of *Aedes aegypti* [21]. GM18, which carried no *cry* or *cyt* kind gene, was reported to have no insecticidal activity, while its solubilized inclusion proteins displayed cytotoxic activity against to several cancer cell lines buy not to sheep erythrocytes [20] — a phenotype consistent with a parasporin-like rather than an insecticidal repertoire, and one that makes its single named App4 locus worth following up. IB84, which carried no classical

insecticidal locus, produced a small parasporal crystal whose proteins were not toxic to the insects tested but it showed biocidal activity against some cancer cell line [22].

The concordance should not be overstated. These are presence or absence observations from preliminary assays reported in announcement format; no dose–response relationship, no LC50 and no expression data exist for any of the three strains, and none were generated in this work. What the agreement supports is the direction of the inference — that gene content and not species name is what tracks the phenotype — and not any quantitative prediction of potency.

Genomic relatedness does not predict toxin repertoire

Across the 190 pairs, repertoire overlap measured as the Jaccard index over toxin family sets was low and essentially flat against ANI below the species threshold, tightening only among near-clonal pairs above 99% ANI (Fig. 5, Additional file 2: Table S3). Knowing that two genomes share 91% or 94% identity conveys no information about which toxins they carry.

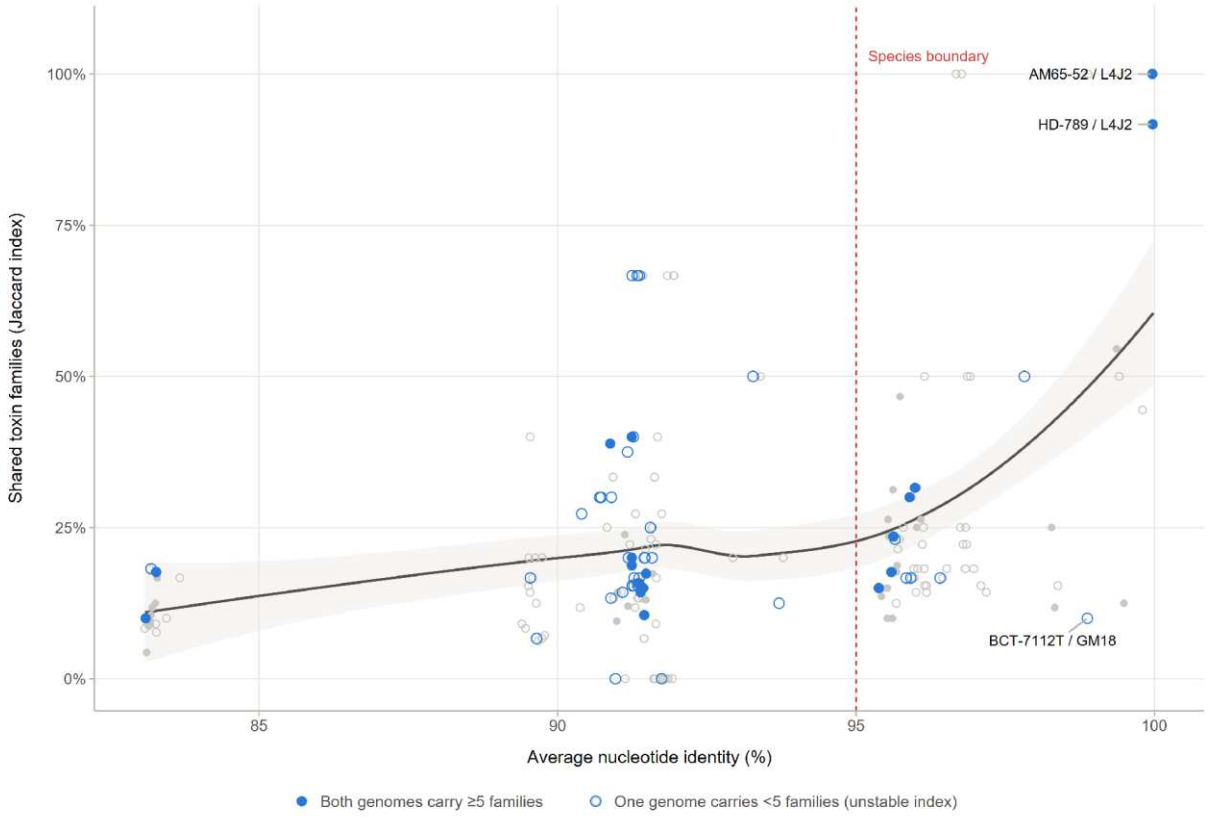


Fig. 5. Toxin-family repertoire overlap versus average nucleotide identity. Each point is one of the 190 possible genome pairs, with average nucleotide identity on the abscissa and the Jaccard index between their toxin and parasporin-like family repertoires on the ordinate. The line is a local regression with a 95% confidence band and the dashed line marks the 95% ANI species threshold; open points mark pairs in which either genome carries fewer than five families, a condition under which the index is unstable and which must not be read as biological similarity. Below the species threshold, repertoire overlap is low and independent of ANI; only among

near-clonal pairs do the two quantities correspond. The two contrasting cases illustrate the effect: L4J2 and the references AM65-52 and HD-789 share 99.97% identity and identical or near-identical repertoires, whereas GM18 and the type strain of its own species, BCT-7112T, share 98.87% identity but almost no toxin content — a consequence of these genes residing mainly on plasmids and mobile elements transmitted independently of the chromosome. The index is computed on family presence and absence, without weighting by locus count or evidence tier. The local regression is descriptive and its confidence band widens where pairs are sparse; it must not be read as prediction.

The two contrasting cases are both drawn from the study strains. L4J2 against AM65-52 and HD-789: 99.96% and 99.97% ANI and, respectively, an identical family set (Jaccard 1.00, twelve families each) and a near-identical one (0.92, eleven families). This is why L4J2 carried the *israelensis* cassette — not functional convergence, but the same lineage still holding the same plasmids. GM18 against BCT-7112T: 98.87% ANI, unambiguously the same species and the same chromosomal backbone, but almost no shared toxin content. The substance of this second contrast is best stated as counts rather than as an index: GM18 carried 18 loci across 10 families, whereas the type strain of its own species carried a single locus in a single family. We note explicitly that the Jaccard value for this pair (0.10) rests on a repertoire of one family in BCT-7112T and is for that reason unstable.

That instability is general and constrains the figure. Family counts per genome ranged from one (BCT-7112T) to 15 (FCC41), and IB84 carried only three, so pairs in which either genome had a small repertoire could reach high Jaccard values for trivial reasons — IB84 against ATCC 14579T scored 0.67 on two shared families, neither of them an insecticidal arsenal. Those pairs are marked as open points in Fig. 5 and are not interpreted as biological similarity. The index is also computed on family presence and absence, without weighting by locus count or evidence tier, so given how few structure-based loci reach named-gene identity, repertoire overlap in that class is considerably more uncertain than in the classical one. Finally, this is a negative result: it demonstrates absence of association within this panel and does not prove that no relationship exists at other sampling densities or taxonomic scales.

The mobilome accounts for the decoupling between taxonomy and gene content

If these genes travel on replicons that are inherited independently of the chromosome, then neither ANI nor a core-genome phylogeny — both of which measure the chromosome — can predict them. Four lines of evidence from this dataset support that account.

The genes were on plasmids. Of the 271 locus entries, 185 (68.3%) resided on plasmids and 86 on chromosomes (Additional file 1: Fig. S9). The two classes behaved differently: classical insecticidal families were markedly plasmid-borne, whereas the structure-based class was radically heterogeneous, ranging from entirely plasmid-borne families to essentially chromosomal ones (Additional file 1: Fig. S10, Fig. S11, Additional file 2: Table S5). Spp1, the most ubiquitous family of the panel, was also among the least plasmid-borne of its class, which is the profile of a conserved chromosomal gene rather than a mobile parasporal determinant. In the three study strains the pattern was unusually clean (Table 2): all 14 of L4J2's classical insecticidal loci were plasmid-borne, as were both of its named Mpp60 loci, and the single locus it carried on a chromosome was an Spp1 call. All three of IB84's loci were plasmid-borne, as

were 16 of GM18's 18, including its named App4 locus. That L4J2's entire mosquitocidal cassette sits on plasmids is precisely how a strain can carry the arsenal without carrying the name.

The plasmid localization was specific to this gene class, not a general property of the panel. Under a Monte Carlo null that redistributes loci across replicons in proportion to replicon length, toxin loci were enriched 4.54-fold on plasmids, at 68.3% observed against 15.0% expected, $p < 0.0001$ (Fig. 6). The test discriminated sharply between categories: CRISPR spacer targets were enriched 7.81-fold (29 of 29 on plasmids), insertion sequences 3.57-fold (769 of 1596), bacteriocins 2.47-fold (23 of 62), defense systems 2.34-fold (82 of 261) and biosynthetic clusters 1.59-fold (56 of 256), all significant, while antimicrobial resistance genes showed no enrichment (1.04-fold, 21 of 139 loci, $p = 0.47$) and virulence factors fell below expectation (0.46-fold, 27 of 429, $p = 1.00$). Six of the eight categories tested were significantly enriched. The contrast between toxins and virulence factors is the most informative result of this test: two gene classes with opposite inheritance patterns in the same genomes, which is what one expects if the insecticidal arsenal is an accessory mobile module superimposed on a stable chromosomal background. Because the test is one-sided and designed to detect enrichment, the virulence result licenses the description "not enriched" rather than "significantly depleted".

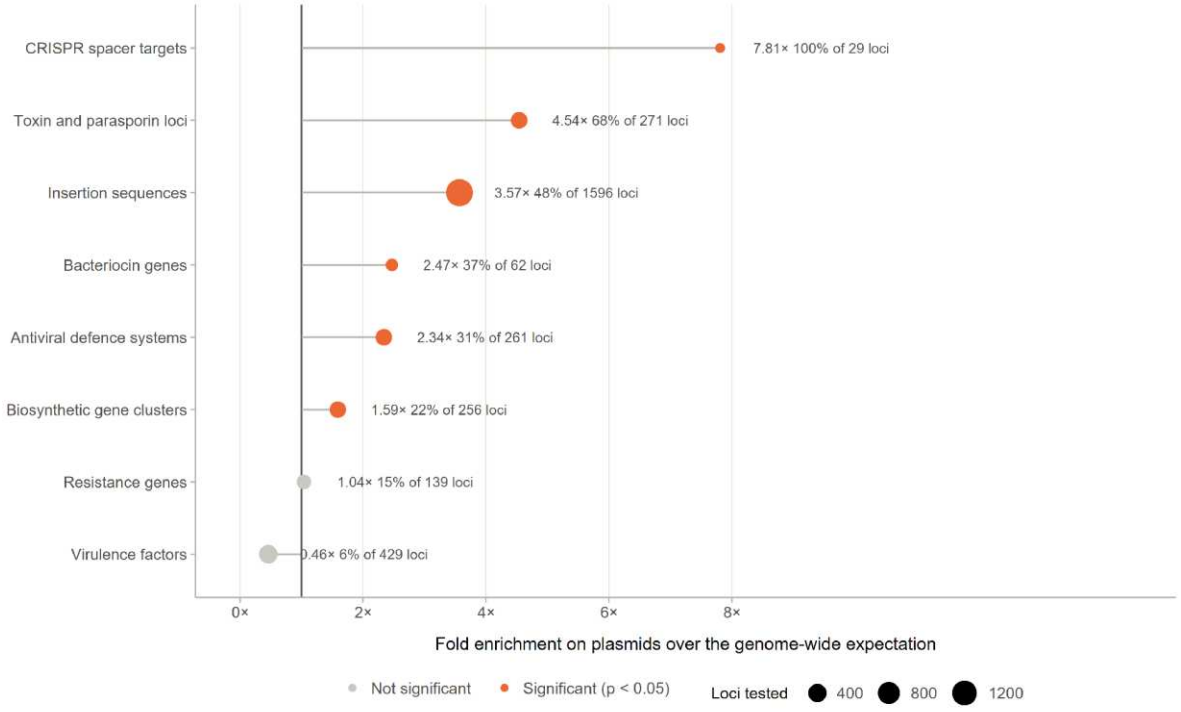


Fig. 6. Plasmid enrichment of genomic feature categories. For each category, the enrichment factor is the observed proportion of loci on plasmids divided by the proportion expected under a null model that redistributes loci at random across replicons in proportion to replicon length, assessed by Monte Carlo randomisation. Point size is proportional to the number of loci tested and colour indicates statistical significance; the vertical line marks absence of enrichment. Toxin and parasporin-like loci are enriched 4.54-fold, with 68.3% of their loci on plasmids against

15.0% expected, whereas virulence factors fall below expectation and resistance genes show no significant enrichment. The null model assumes loci could lie on any replicon in proportion to its size, which is an idealisation. The test is one-sided and designed to detect enrichment, so the virulence value of 0.46-fold licenses the description “not enriched” rather than “significantly depleted”. The test aggregates the twenty genomes and does not distinguish between them.

The loci sat in insertion-sequence-rich neighbourhoods. Toxin loci lay closer to insertion sequences than expected under random placement within the same replicon, and the excess was strongest at the shortest range: 100 loci fell within 1 kb of an insertion sequence against 71.2 expected (a 1.40-fold excess), 146 within 5 kb against 109.2 expected (1.34-fold) and 168 within 10 kb against 131.6 expected (1.28-fold) (Additional file 1: Fig. S12) [65, 66]. Insertion sequences are a well-documented driver of gene inactivation and genome rearrangement in bacteria [86]. Specificity, rather than the excess itself, is the finding: at the 10 kb window, virulence factors (66 against 59.5 expected, 1.11-fold), resistance genes (50 against 40.2, 1.24-fold) and biosynthetic clusters (166 against 159.1, 1.04-fold) depart far less from expectation under the same test. Bacteriocin genes show a sharp excess at the shortest range (18 against 5.6 expected at 1 kb, 3.21-fold) that decays by 10 kb (26 against 20.4, 1.27-fold), but on a much smaller base than toxins, which makes that estimate less stable; we report it rather than presenting toxins as the only associated class. Four caveats bound this analysis. It is restricted to the complete genomes of the panel, because contig edges prevent reliable distance measurement in drafts, so the three study strains do not contribute to it directly. The three window sizes are nested and therefore not independent. We observed and expected counts and their ratios, and do not attach significance levels to them, because the per-category significance of this test was not recomputed in the final analysis. And association does not establish direction: insertion sequences may insert preferentially near these loci, both may favor the same genomic regions, or the loci may have moved into regions already rich in them.

The plasmids demonstrably circulated, and the bacteria retained the record. Only five of the 20 genomes carried a complete CRISPR-Cas system, all of subtype I-C (GM18, DSM 12442T, HD-1, HD133 and HS18-1), so only five could accumulate spacer memory at all; the remaining fifteen are absent from Additional file 1: Fig. S13 for that reason and not for lack of encounters. Among those five, 45 confident spacer matches against mobile elements were recovered, 41 of them against elements carried by *other* genomes of the panel, and of the targets identified, 29 were plasmids rather than prophages: what these lineages have learned to recognize is predominantly plasmid DNA [74]. GM18 was the only study strain with such memory, contributing six spacer matches against foreign elements, and it was also the genome with the largest number of defense systems in the panel (25, of which one was plasmid-resident) (Additional file 1: Fig. S14, Fig. S15). Spacer targets were the most plasmid-enriched category in the entire analysis (7.81-fold, every target on a plasmid). Absence of memory does not imply absence of encounter, and a spacer match dates and directs nothing; what the analysis establishes is that plasmid-borne elements move between lineages of this group. All three study strains additionally carried conjugative plasmids of at least 10 kb — three in GM18, three in IB84 and two in L4J2 — so the mobility inferred from spacer memory has a plausible mechanism in these genomes.

Biosafety-relevant gene content, and the limits of reading it

The virulence repertoire was largely uniform across the panel and was dominated by determinants shared throughout *B. cereus sensu lato* rather than by strain-specific factors (Additional file 1: Fig. S16). Confident virulence-factor clusters ranged from three to nine per genome, and the study strains sat inside that range: GM18 carried five, the fewest of the three, and IB84 and L4J2 eight each. The non-haemolytic enterotoxin NHE, the haemolysin BL complex HBL and the immune inhibitor InhA were present in all 20 genomes, including the most divergent. These are the determinants associated with *B. cereus* foodborne illness [1, 2], and their presence across much of the panel — including reference strains in long-standing commercial use — is the context in which any equivalent finding for the study strains must be read: it does not distinguish a candidate strain from formulations deployed for decades. Two explicit safety checks returned negative results in all 20 genomes: no complete cereulide synthetase operon was detected, with only an orphan *cesH* hit in BMB171 and no cereulide cluster among the 256 biosynthetic regions [83], and no anthrax toxin genes were found.

Antimicrobial resistance content was dominated by the intrinsic chromosomal β-lactamases BlaI, BlaII and BlaIII, present across essentially the whole panel and part of the *B. cereus* group gene pool rather than acquired resistance (Additional file 1: Fig. S17). Panel-wide, 139 resistance loci were called and only 21 of them lay on plasmids, which is no more than expected from replicon length (1.04-fold, $p = 0.47$). L4J2 carried the highest number of resistance loci in the panel (11, four of them plasmid-located), IB84 nine (two plasmid-located) and GM18 seven (one plasmid-located); the reference strains in established commercial use carry comparable repertoires, so none of the three is anomalous within the group. A *vanR*/*vanS* pair was detected in IB84 and was assigned to the chromosome. Detection of a resistance gene is in any case not evidence of resistance, which requires susceptibility testing, and none was performed here.

Secondary metabolite capacity was highest in L4J2, with 17 biosynthetic regions, the maximum of the panel, out of 256 across all 20 genomes (Additional file 1: Fig. S18). This is largely a property of the clade rather than of L4J2 specifically, though it remains a real and reportable feature of the strain; it should also be read alongside L4J2's contamination estimate, since redundant sequence would inflate a region count. GM18 and IB84 each carried 13 regions and sat in the lower-middle of the panel. The bacteriocin repertoire was modest and uniform, with no strain standing out: the confident calls were Lactococcin 972 and UviB in GM18, UviB in IB84, and UviB and uberolysin in L4J2 (Additional file 1: Fig. S19). Petrobactin and bacillibactin siderophore systems were detected concordantly by two independent routes. A predicted biosynthetic region does not imply production of the metabolite, which depends on expression and growth conditions.

The mobile compartment, strain by strain

The three strains differed markedly in mobile-element content, and the within-genome distribution was more informative than the totals. Plasmid sequence accounted for 20.7% of GM18, 22.7% of IB84 and 26.3% of L4J2; the panel maximum was HD133 at 28.1%, so none of the three was exceptional in this respect. Insertion-sequence density, however, was strongly partitioned by replicon in all three: GM18 carried 2.26 IS per Mb on its chromosome against 57.47 per Mb on its plasmids, IB84 1.97 against 30.49, and L4J2 2.13 against 22.25. Across the panel 769 of 1596 insertion sequences lay on plasmids, a 3.57-fold enrichment (Additional file 1:

Fig. S20). The plasmid compartment of these genomes was where both the insecticidal loci and the transposition machinery were concentrated, which is the replicon-level statement of the association reported above.

Integrated prophages follow the same asymmetry between the study strains but not the same ranking: GM18 carried two integrated prophages occupying 0.87% of its genome and IB84 one occupying 0.31%, whereas L4J2 carried ten occupying 3.98%, three of them on plasmids (Additional file 1: Fig. S21). L4J2's burden was among the highest of the panel, though ST7 and A1100 each carried eleven. Defence systems were distributed in the opposite direction: GM18 carried 25, the panel maximum, against 16 in IB84 and 18 in L4J2, and only one of GM18's was plasmid-resident. Panel-wide, 82 of 319 defence systems (25.7%) were plasmid-resident. GM18 was thus the one study strain that combined an active CRISPR-Cas system, the largest defence-system complement of the panel and a plasmid compartment dense in insertion sequences, without carrying any classical insecticidal locus.

Pangenome context

The pangenome of the 20 genomes comprised 20 731 gene families, of which 1934 were core (present in 99–100% of strains), 1647 soft core, 4954 shell and 12 196 cloud [43]. Variation between strains resided almost entirely in the accessory and strain-specific compartments (Additional file 1: Fig. S22), which is the expected signature of an open pangenome. eggNOG-mapper assigned a function to 16 257 of the 19 909 gene-cluster representatives (81.7%) [55, 56]. The strain-specific compartment was small in all three study strains and smallest by far in L4J2 (276 genes in GM18, 240 in IB84 and 12 in L4J2), the last reflecting the near-clonal reference genomes that flank it in the panel rather than any property of the strain itself (Additional file 1: Fig. S23). Poorly characterized categories dominated the strain-specific compartment in every genome, and among the categories with defined function, replication, recombination and repair were over-represented relative to a typical genome (Additional file 2: Table S6) — consistent with a compartment enriched in transposases, integrases and phage genes (Additional file 1: Fig. S24). The three study strains showed no anomalous functional profile relative to the references. That toxin genes reside in this accessory, mobile compartment rather than in the core is the pangenome-level statement of the same result: the arsenal is not part of what the species shares.

Discussion

Taxonomic identity and insecticidal gene content behaved as independent properties in this panel. The first is a property of the chromosome, recovered concordantly by average nucleotide identity, digital DNA–DNA hybridization and a fully-resolved core-genome phylogeny. The second sits on plasmids, in neighbourhoods dense in insertion sequences. In a compartment that the CRISPR memory of these same genomes showed to be circulating between lineages. Below the species boundary, neither predicts the other.

The three strains examined here make the point in a form that a statistical statement cannot. L4J2 is deposited as *B. cereus* group and carried the complete canonical mosquitocidal cassette at named-gene identity, in the same five families and the same seven proteins as the two *israelensis* references, entirely plasmid-borne: a screen guided by the species name would have discarded it.

GM18 is unambiguously *B. toyonensis*, sharing 98.87% ANI, 90.5% dDDH and a 100%-supported sister relationship with the type strain of its species. Yet it carried no *cry* or *cyt* gene and held 18 toxin *loci* across 10 families, where the holotype strain maintained one locus in one family: identifying a strain to species conveys no information about its arsenal, even when the identification is correct and precise. IB84 shows that absence is equally informative — it came from sampling directed at *B. thuringiensis* and carried no classical insecticidal locus whatsoever, making it a clean internal negative control and a non-toxigenic close relative worth having in the panel. It is not alone in that condition: BMB171 and Bt407, both deposited as *B. thuringiensis*, were equally empty of classical insecticidal loci, which is the same dissociation seen from the opposite direction. The operational recommendation follows directly: in prospecting for biocontrol agents, species identification is not a substitute for direct profiling of pesticidal gene content, and the profiling must report sequence-identity tiers and reading-frame integrity rather than locus counts alone.

In all three strains, it was the gene content, and not the species name under which they had been identified, that tracked the activity reported for them (Additional file 2: Table S9). That agreement is the only external test of the inference available here, and it is qualitative: it supports the direction of the argument and has no quantitative prediction. It also gives the GM18 result a concrete follow-up. Parasporal proteins from isolates of this same collection are cytotoxic to the human leukemia line MOLT-4 [87], so GM18's single named App4 locus — in a *B. toyonensis*, a species absent from the parasporin literature — is the first candidate worth testing against that panel of cell lines.

These three cases also show what genome sequencing contributes to a strain collection assembled on phenotype. All three were recovered by a selective procedure designed to isolate *B. thuringiensis* [18] and were maintained and reported under that name on the classical criteria: formation of a spore-associated parasporal crystal in all three, and flagellar serotyping in the case of GM18, which was assigned a new H serotype [20, 22, 21]. Each of the three was observed to produce a parasporal inclusion, so that character was satisfied in every case. Yet whole-genome comparison placed GM18 in *B. toyonensis* and IB84 in *B. albus*, neither of which the phenotypic criteria separate from *B. thuringiensis*, and it showed L4J2 — the one strain not reported under the *thuringiensis* name — to be the one carrying the complete mosquitocidal cassette. Parasporal crystal formation is thus a capacity shared across several species of this group rather than a diagnostic character for one, and a collection curated on that character will accumulate a mixture of species under a single label. The mixture becomes visible only when the chromosomes are sequenced. This is no criticism of the original identifications, which were sound by the criteria available when they were made; it is the reason a collection of applied interest repays genomic re-examination even when no new taxon is proposed, and the reason the two properties that matter for biocontrol — what the strain is and what it carries — have to be established separately.

A nomenclatural corollary strengthens the same argument from the taxonomic side. Seven of the ten reference strains deposited as *B. thuringiensis* were assigned to *B. cereus* on genomic criteria, including HD-1 and HD73, the subsp. *kurstaki* strains that have underpinned commercial formulations for decades. This is not a defect of the data; it is the expected consequence of classifying on a chromosome a group whose defining applied phenotype lives on plasmids. The practical implication is that "*B. thuringiensis*" denotes different things in the applied literature and in genome taxonomy, and that any study using the name should state which system it means

— a convention we have followed explicitly throughout. We note that our panel does not contain the type strain of *B. thuringiensis*, so these observations concern the reference strains examined and not the taxon as a whole.

The mechanism we document also carries a less comfortable implication for applied use. The mobility that spreads a toxin cassette between lineages is the same mobility that degrades it within one. Toxin loci were enriched 4.54-fold on plasmids and lay closer to insertion sequences than expected at every window tested, while virulence factors — not plasmid-enriched, and no closer to insertion sequences than expected — were not; and L4J2, which carried ten integrated prophages occupying 3.98% of its genome, retained an intact reading frame in only 11 of its 24 toxin loci, including pseudogenes confirmed independently by two annotation pipelines. Carrying a cassette and expressing it are different claims, and at no point does this work claim insecticidal activity. For a candidate biopesticide, this bears on product stability, as a plasmid-borne determinant can be lost during culture, and on regulatory assessment, since a mobilizable replicon carrying accessory genes may transfer. The specificity of the association is what makes it interpretable: under the same test, virulence genes, resistance genes and biosynthetic clusters depart far less from expectation than toxin loci do, so this is not a general property of these genomes. Three constraints apply. The test establishes association and not direction. It rests on the 14 closed genomes of the panel and therefore not on the three strains reported here, whose contig edges prevent reliable distance measurement. And we report observed and expected counts with their ratios but attach no significance levels to them, because the per-category significance of this test was not recomputed in the final analysis.

One methodological result has interest beyond this group. A set of 20 vertically inherited marker genes placed each of the three study strains correctly, yet it did not recover the topology that the complete core-genome alignment supports at near-universal 100% bootstrap: the two trees shared only four internal nodes and differed at 26 of a possible 34 bipartitions, and the marker inference declared three genome pairs exactly identical. The discrepancies concentrate where divergence is lowest, which points to insufficient resolution rather than an alternative history. A direct test of that interpretation arose in the course of this work. When one assembly of the panel was replaced by a more contiguous version of the same strain, the 3399-gene core-genome topology was unchanged (Robinson–Foulds distance 0), whereas the 20-gene topology moved by a Robinson–Foulds distance of 16 and relocated that strain from a position outside its own species to its correct sister relationship with its type strain at 100% bootstrap support. A reduced marker scheme is thus sensitive to assembly input at a magnitude that the core alignment absorbs entirely. Reduced marker schemes are under-powered inside the *B. cereus* group, and a marker-based tree should not be presented as corroborating a core-genome tree in a panel containing many pairs above 95% ANI.

Reading the biosafety-relevant content requires the same care. The enterotoxin determinants associated with *B. cereus* foodborne illness are present across most of the panel, including in reference strains in long-standing commercial use, so their presence does not distinguish a candidate strain from deployed formulations; two explicit checks — for the cereulide synthetase operon and for anthrax toxin genes — were negative in all 20 genomes. Resistance content is dominated by intrinsic chromosomal β-lactamases of the group rather than acquired determinants, and panel-wide only 21 of 139 resistance loci lay on plasmids, with no significant plasmid enrichment for that category. One observation warrants explicit declaration: a *vanR*/*vanS*

pair was detected in IB84 and assigned to the chromosome. Detection of a resistance or virulence gene is not evidence of the phenotype, which requires susceptibility testing.

Three corrections to the published record follow from this analysis, and we state them explicitly rather than leaving them to be inferred. First, the genome announcement for GM18 reported two *cry*-like coding sequences in that strain [20]. The present analysis, which screens the deposited assembly against both the classical and the structure-based nomenclature with identity tiers declared, finds no *cry* or *cyt* gene in GM18 at any identity threshold; its only classical-panel *loci* were three distant Vpb1 matches below 45% identity. The earlier report should be read as a database annotation that the tier-resolved screen does not support, and GM18 should not be cited as a *cry*-bearing strain. Second, an earlier version of our own analysis reported a *bla*TEM-116 allele in GM18. That call lay on an isolated 888 bp contig which the NCBI contamination screen subsequently removed from the deposited assembly; it is absent from the sequence analyzed here and was contamination rather than a GM18 gene. We withdraw it, and we note that analyzing the deposited assembly rather than a local re-assembly is what surfaced the error.

Third, the genome announcement for L4J2 tabulates a genome size of 7 649 903 bp at 35.23 mol% G+C [21], whereas the assembly deposited under GCA_042535225.1 — the sequence analyzed here — comprises 7 016 826 bp in 799 contigs at 34.96 mol% G+C. The difference arises after submission: NCBI's contamination screen removed whole contigs from the submitted assembly without splitting or trimming any of those retained, and the removed material accounts for both the missing length and the shift in G+C content. The announcement's figures also pair a contig count filtered at 500 bp with an unfiltered total length, so they are not mutually consistent. We therefore cite the deposited record throughout and treat the announcement's genome size as an erratum. This is a general point about provenance rather than a criticism of that report: when the analyzed sequence is a local assembly, post-submission curation can silently separate the published numbers from the public record, which is the reason we analyzed deposited assemblies for all three strains.

Limitations

Everything reported here is genomic potential. There is no genetic expression data for any of the three strains, and the only phenotypic information available is the qualitative characterization in their genome announcements; any phenotypic facts generated in this work. The sequence-identity tiers we report are thresholds against a reference database.

Two assembly-level constraints bound specific numbers. L4J2's contamination exceeds the 5% high-quality ceiling on both independent estimates (CheckM2 5.97%, PGAP CheckM 9.07%), and it is the most fragmented assembly of the panel at 799 contigs, so its genome size and every count that scales with it — biosynthetic regions, prophages, locus totals — are upper bounds pending a closed assembly. GM18's intact-ORF fraction, by contrast, is not an assembly artefact: re-examining the strain on an assembly of 156 rather than 1193 contigs left the pseudogene load unchanged, so long reads would not recover the missing frames, although the assignment of individual loci to failure modes remains provisional.

Several measures are lower bounds or restricted in scope. Plasmid reconstruction from drafts is incomplete, so plasmid counts are minimal, and mobility is predicted from genetic determinants rather than demonstrated; replicon assignment is declared only for the 14 closed genomes. Most

structure-based parasporin-like assignments fall below 45% identity; the reported numbers represent candidates rather than confirmed hits, and the ubiquitous Spp1 family was excluded from presence/absence interpretation. The Jaccard index is unstable for small repertoires, and pairs with fewer than five families are marked and not interpreted. The insertion-sequence proximity analysis excludes the draft genomes and is reported as observed and expected counts without significance levels, since the per-category test was not recomputed in the final analysis.

Finally, there are two gaps in the source data. The two nomenclature panels yield 271 locus entries against 243 distinct loci after merging, and the 16 dual-classified loci do not account for the difference, so we report both totals and interpret neither. The exact sampling dates of the three isolates are not recorded in the collection material available to us.

Conclusions

A twenty-genome panel built with deliberate internal controls — one strain with the arsenal but not the name, one with the name but not the arsenal, and one with neither — allows the decoupling of taxonomy from applied function in the *Bacillus cereus* group to be demonstrated quantitatively rather than argued from anecdote. Species identity is recovered from the chromosome, concordantly and robustly; the insecticidal arsenal travels on plasmids, in insertion-sequence-rich neighbourhoods, and is gained and lost independently of that chromosome. The consequence for anyone screening environmental isolates for biocontrol candidates is concrete: identify the strain but profile its genes.

List of abbreviations

ANI, average nucleotide identity; BGC, biosynthetic gene cluster; BPPRC, Bacterial Pesticidal Protein Resource Center; dDDH, digital DNA–DNA hybridization; GTDB, Genome Taxonomy Database; IS, insertion sequence; UFBoot, ultrafast bootstrap.

Declarations

Ethics approval and consent to participate

Not applicable. This study involved no human subjects, no animal experiments, no clinical samples and no clinical trial; it is a comparative analysis of previously published bacterial genome sequences.

Consent for publication

Not applicable.

Availability of data and materials

All genome sequence data underlying this study are deposited in public databases and were publicly available before submission. The assemblies analyzed are the deposited records themselves, so each accession below identifies the exact sequence used.

1. Genome assemblies of the three study strains. GM18, GenBank assembly GCA_015620955.1 (RefSeq GCF_015620955.1; WGS project JADNYS01; BioSample SAMN16746850; BioProject PRJNA676032). IB84, GCA_015620945.1 (RefSeq GCF_015620945.1; WGS JADNYT01; BioSample SAMN16746851; BioProject PRJNA676033). L4J2, GCA_042535225.1 (RefSeq GCF_042535225.1; WGS JBFBPH01; BioSample SAMN42206399; BioProject PRJNA1130484).
2. Raw sequencing reads, used to validate each deposited assembly by read mapping, are in the Sequence Read Archive: GM18, run SRR13402414 (experiment SRX9821488, study SRP301246); IB84, SRR13427799 (SRX9842884, SRP301583); L4J2, SRR29744432 (SRX25245429).
3. Prior reports of the three genomes. The assemblies and the preliminary phenotypic characterizations were first described in three genome announcements: GM18, doi:10.1128/mra.00008-21; IB84, doi:10.1128/mra.00274-21; L4J2, doi:10.1128/mra.01180-24. Where the figures tabulated in those reports differ from the deposited records, the deposited records are cited here, and the discrepancy is addressed in the Discussion.
4. The 17 comparison genomes are publicly available; their assembly accessions are listed in Additional file 2: Table S7 and repeated in the Methods.
5. Underlying data for every figure. Additional file 2: Tables S1–S6 and S9 contain the per-pair, per-locus, per-family and per-strain values behind each figure of the main text and are supplied as supplementary files.
6. Software. Every tool is listed with version, purpose, non-default parameters, and reference-database version in Additional file 2: Table S8, and each is cited in the Methods with the DOI of its primary publication.
7. Code. No custom algorithm was developed for this study. The analysis scripts that generated the figures and tables of this work are archived at Zenodo under doi:10.5281/zenodo.23322269 (https://doi.org/10.5281/zenodo.23322269).
8. Phenotypic data. No phenotypic, bioassay, or gene-expression data were generated in this study. The qualitative phenotypes compared with gene content in Additional file 2: Table S9 are taken from the three genome announcements cited in item 3.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was partially financed by Universidad Autónoma de Nuevo León. Y.H. is supported by a fellowship from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Authors' contributions

C.A.F.-C., conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing - original draft. Y.H., methodology, validation, writing - review and editing. A.S.L., supervision, resources, writing - review and editing. E.L.-L., A.S.-R., J.L.G.-G., S.A.M.M., M.E.S. and L.J.G.-W., investigation, resources, writing - review and editing. B.P.-

A., conceptualization, project administration, supervision, resources, writing - review and editing. All authors read and approved the final manuscript.

Acknowledgements

We thank the Instituto de Biotecnología of the Facultad de Ciencias Biológicas, Universidad Autónoma de Nuevo León, for access to its facilities and for maintaining the Bacillus thuringiensis strain collection from which the three strains examined here were drawn.

References

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

2. Ehling-Schulz M, Lereclus D, Koehler TM. The Bacillus cereus Group: Bacillus Species with Pathogenic Potential. Microbiol Spectr. 2019;7. doi:10.1128/microbiolspec.gpp3-0032-2018

3. Helgason E, Økstad OA, Caugant DA, Johansen HA, Fouet A, et al. Bacillus anthracis, Bacillus cereus, and Bacillus thuringiensis—One Species on the Basis of Genetic Evidence. Appl Environ Microbiol. 2000;66:2627-2630. doi:10.1128/aem.66.6.2627-2630.2000

4. Liu Y, Lai Q, Göker M, Meier-Kolthoff JP, Wang M, et al. Genomic insights into the taxonomic status of the Bacillus cereus group. Sci Rep. 2015;5. doi:10.1038/srep14082

5. Carroll LM, Wiedmann M, Kovac J. Proposal of a Taxonomic Nomenclature for the Bacillus cereus Group Which Reconciles Genomic Definitions of Bacterial Species with Clinical and Industrial Phenotypes. mBio. 2020;11. doi:10.1128/mbio.00034-20

6. Liu Y, Du J, Lai Q, Zeng R, Ye D, et al. Proposal of nine novel species of the Bacillus cereus group. International Journal of Systematic and Evolutionary Microbiology. 2017;67:2499-2508. doi:10.1099/ijsem.0.001821

7. González JM, Brown BJ, Carlton BC. Transfer of Bacillus thuringiensis plasmids coding for delta-endotoxin among strains of B. thuringiensis and B. cereus. Proc. Natl. Acad. Sci. U.S.A. 1982;79:6951-6955. doi:10.1073/pnas.79.22.6951

8. Berry C, O'Neil S, Ben-Dov E, Jones AF, Murphy L, et al. Complete Sequence and Organization of pBtoxis, the Toxin-Coding Plasmid of Bacillus thuringiensis subsp. israelensis. Appl Environ Microbiol. 2002;68:5082-5095. doi:10.1128/aem.68.10.5082-5095.2002

9. Crickmore N, Zeigler DR, Feitelson J, Schnepf E, Van Rie J, et al. Revision of the Nomenclature for the Bacillus thuringiensis Pesticidal Crystal Proteins. Microbiol Mol Biol Rev. 1998;62:807-813. doi:10.1128/mmbr.62.3.807-813.1998

10. Crickmore N, Berry C, Panneerselvam S, Mishra R, Connor TR, et al. A structure-based nomenclature for Bacillus thuringiensis and other bacteria-derived pesticidal proteins. Journal of Invertebrate Pathology. 2021;186:107438. doi:10.1016/j.jip.2020.107438

11. Bravo A, Likitvivatanavong S, Gill SS, Soberón M. Bacillus thuringiensis: A story of a successful bioinsecticide. Insect Biochemistry and Molecular Biology. 2011;41:423-431. doi:10.1016/j.ibmb.2011.02.006

12. Sanahuja G, Banakar R, Twyman RM, Capell T, Christou P. Bacillus thuringiensis: a century of research, development and commercial applications. Plant Biotechnology Journal. 2011;9:283-300. doi:10.1111/j.1467-7652.2011.00595.x

13. Ben-Dov E. Bacillus thuringiensis subsp. israelensis and Its Dipteran-Specific Toxins. Toxins. 2014;6:1222-1243. doi:10.3390/toxins6041222

14. Lacey LA. BACILLUS THURINGIENSIS SEROVARIETY ISRAELENSIS AND BACILLUS SPHAERICUS FOR MOSQUITO CONTROL. Journal of the American Mosquito Control Association. 2007;23:133-163. doi:10.2987/8756-971x(2007)23[133:btsiab]2.0.co;2

15. Mizuki E, Ohba M, Akao T, Yamashita S, Saitoh H, et al. Unique activity associated with non-insecticidal Bacillus thuringiensis parasporal inclusions: in vitro cell-killing action on human cancer cells. J Appl Microbiol. 1999;86:477-486. doi:10.1046/j.1365-2672.1999.00692.x

16. Mizuki E, Park YS, Saitoh H, Yamashita S, Akao T, et al. Parasporin, a Human Leukemic Cell-Recognizing Parasporal Protein of Bacillus thuringiensis. Clin Diagn Lab Immunol. 2000;7:625-634. doi:10.1128/cdli.7.4.625-634.2000

17. Ekino K, Okumura S, Ishikawa T, Kitada S, Saitoh H, et al. Cloning and Characterization of a Unique Cytotoxic Protein Parasporin-5 Produced by Bacillus thuringiensis A1100 Strain. Toxins. 2014;6:1882-1895. doi:10.3390/toxins6061882

18. Travers RS, Martin PAW, Reichelderfer CF. Selective Process for Efficient Isolation of Soil Bacillus spp. Appl Environ Microbiol. 1987;53:1263-1266. doi:10.1128/aem.53.6.1263-1266.1987

19. Rodríguez-Padilla C, Galán-Wong L, de Barjac H, Román-Calderón E, Tamez-Guerra R, et al. Bacillus thuringiensis subspecies *neoleonensis* serotype H-24, a new subspecies which produces a triangular crystal. J Invertebr Pathol. 1990;56:280-282. doi:10.1016/0022-2011(90)90111-I

20. Moreno-Peña CJ, García-García JH, Hernández-Vásquez CI, Galán-Wong LJ, Pereyra-Alférez B. Draft Genome Sequence of *Bacillus toyonensis* Strain GM18, Isolated from Agricultural Soil. Microbiol Resour Announc. 2021;10. doi:10.1128/mra.00008-21

21. Rangel-Heredia J, Flores-Cruz CA, Hernández-Vásquez CI, Martínez-Morales SA, Galán-Wong LJ, et al. *Bacillus cereus* strain L4J2, isolated from Mexican soil, carries multiple genes for biotechnological application. Microbiol Resour Announc. 2025;14. doi:10.1128/mra.01180-24

22. García-García JH, Moreno-Peña CJ, Hernández-Vásquez CI, Galán-Wong LJ, Pereyra-Alférez B. Draft Genome Sequence of Bacillus albus Strain IB84, Isolated from Mexican Soil. Microbiol Resour Announc. 2021;10. doi:10.1128/mra.00274-21

23. Miller RA, Beno SM, Kent DJ, Carroll LM, Martin NH, et al. Bacillus wiedmannii sp. nov., a psychrotolerant and cytotoxic Bacillus cereus group species isolated from dairy foods and dairy environments. International Journal of Systematic and Evolutionary Microbiology. 2016;66:4744-4753. doi:10.1099/ijsem.0.001421

24. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil PA, et al. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Research. 2021;50:D785–D794. doi:10.1093/nar/gkab776

25. Parks DH, Chuvochina M, Chaumeil PA, Rinke C, Mussig AJ, et al. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol. 2020;38:1079-1086. doi:10.1038/s41587-020-0501-8

26. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072-1075. doi:10.1093/bioinformatics/btt086

27. Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. Nat Methods. 2023;20:1203-1212. doi:10.1038/s41592-023-01940-w

28. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Molecular Biology and Evolution. 2021;38:4647-4654. doi:10.1093/molbev/msab199

29. Vasimuddin M, Misra S, Li H, Aluru S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS) 2019. doi:10.1109/ipdps.2019.00041

30. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10. doi:10.1093/gigascience/giab008

31. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32:3047-3048. doi:10.1093/bioinformatics/btw354

32. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, et al. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microbial Genomics. 2021;7. doi:10.1099/mgen.0.000685

33. Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, et al. RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. Nucleic Acids Research. 2020;49:D1020–D1028. doi:10.1093/nar/gkaa1105

34. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics. 2022;38:5315-5316. doi:10.1093/bioinformatics/btac672

35. Shaw J, Yu YW. Fast and robust metagenomic sequence comparison through sparse chaining with skani. Nat Methods. 2023;20:1661-1665. doi:10.1038/s41592-023-02018-3

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

37. Meier-Kolthoff JP, Auch AF, Klenk HP, Göker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics. 2013;14. doi:10.1186/1471-2105-14-60

38. Meier-Kolthoff JP, Göker M. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. Nat Commun. 2019;10. doi:10.1038/s41467-019-10210-3

39. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. U.S.A. 2009;106:19126-19131. doi:10.1073/pnas.0906412106

40. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. U.S.A. 2005;102:2567-2572. doi:10.1073/pnas.0409727102

41. Wayne LG, Moore WEC, Stackebrandt E, Kandler O, Colwell RR, et al. Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. International Journal of Systematic and Evolutionary Microbiology. 1987;37:463-464. doi:10.1099/00207713-37-4-463

42. Chun J, Oren A, Ventosa A, Christensen H, Arahal DR, et al. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. International Journal of Systematic and Evolutionary Microbiology. 2018;68:461-466. doi:10.1099/ijsem.0.002516

43. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol. 2020;21. doi:10.1186/s13059-020-02090-4

44. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution. 2020;37:1530-1534. doi:10.1093/molbev/msaa015

45. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14:587-589. doi:10.1038/nmeth.4285

46. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution. 2017;35:518-522. doi:10.1093/molbev/msx281

47. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11. doi:10.1186/1471-2105-11-119

48. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011;7:e1002195. doi:10.1371/journal.pcbi.1002195

49. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004;32:1792-1797. doi:10.1093/nar/gkh340

50. Talavera G, Castresana J. Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. Systematic Biology. 2007;56:564-577. doi:10.1080/10635150701472164

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

52. Robinson D, Foulds L. Comparison of phylogenetic trees. Mathematical Biosciences. 1981;53:131-147. doi:10.1016/0025-5564(81)90043-2

53. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2018;35:526-528. doi:10.1093/bioinformatics/bty633

54. Chaudhari NM, Gupta VK, Dutta C. BPGA- an ultra-fast pan-genome analysis pipeline. Sci Rep. 2016;6. doi:10.1038/srep24373

55. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Molecular Biology and Evolution. 2021;38:5825-5829. doi:10.1093/molbev/msab293

56. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research. 2018;47:D309–D314. doi:10.1093/nar/gky1085

57. Liu H, Zheng J, Bo D, Yu Y, Ye W, et al. BtToxin_Digger: a comprehensive and high-throughput pipeline for mining toxin protein genes from Bacillus thuringiensis. Bioinformatics. 2021;38:250-251. doi:10.1093/bioinformatics/btab506

58. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10. doi:10.1186/1471-2105-10-421

59. van der Graaf-van Bloois L, Wagenaar JA, Zomer AL. RFPlasmid: predicting plasmid sequences from short-read assembly data using machine learning. Microbial Genomics. 2021;7. doi:10.1099/mgen.0.000683

60. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. Microbial Genomics. 2018;4. doi:10.1099/mgen.0.000206

61. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, et al. In Silico Detection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing. Antimicrob Agents Chemother. 2014;58:3895-3903. doi:10.1128/aac.02412-14

62. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, et al. Identification of mobile genetic elements with geNomad. Nat Biotechnol. 2023;42:1303-1312. doi:10.1038/s41587-023-01953-y

63. Akhter S, Aziz RK, Edwards RA. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. Nucleic Acids Research. 2012;40:e126–e126. doi:10.1093/nar/gks406

64. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol. 2020;39:578-585. doi:10.1038/s41587-020-00774-7

65. Xie Z, Tang H. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. Bioinformatics. 2017;33:3340-3347. doi:10.1093/bioinformatics/btx433

66. Siguier P. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Research. 2006;34:D32–D36. doi:10.1093/nar/gkj014

67. Ross K, Varani AM, Snesrud E, Huang H, Alvarenga DO, et al. TnCentral: a Prokaryotic Transposable Element Database and Web Portal for Transposon Analysis. mBio. 2021;12. doi:10.1128/mbio.02060-21

68. Gilchrist CLM, Chooi YH. clinker & clustermap.js: automatic generation of gene cluster comparison figures. Bioinformatics. 2021;37:2473-2475. doi:10.1093/bioinformatics/btab007

69. Tesson F, Hervé A, Mordret E, Touchon M, d'Humières C, et al. Systematic and quantitative view of the antiviral arsenal of prokaryotes. Nat Commun. 2022;13. doi:10.1038/s41467-022-30269-9

70. Payne LJ, Meaden S, Mestre MR, Palmer C, Toro N, et al. PADLOC: a web server for the identification of antiviral defence systems in microbial genomes. Nucleic Acids Research. 2022;50:W541–W550. doi:10.1093/nar/gkac400

71. Russel J, Pinilla-Redondo R, Mayo-Muñoz D, Shah SA, Sørensen SJ. CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. The CRISPR Journal. 2020;3:462-469. doi:10.1089/crispr.2020.0059

72. Biswas A, Staals RH, Morales SE, Fineran PC, Brown CM. CRISPRDetect: A flexible algorithm to define CRISPR arrays. BMC Genomics. 2016;17. doi:10.1186/s12864-016-2627-0

73. Biswas A, Gagnon JN, Brouns SJ, Fineran PC, Brown CM. CRISPRTarget. RNA Biology. 2013;10:817-827. doi:10.4161/rna.24046

74. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, et al. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science. 2007;315:1709-1712. doi:10.1126/science.1138140

75. Liu B, Zheng D, Zhou S, Chen L, Yang J. VFDB 2022: a general classification scheme for bacterial virulence factors. Nucleic Acids Research. 2021;50:D912–D917. doi:10.1093/nar/gkab1107

76. Alcock BP, Huynh W, Chalil R, Smith KW, Raphenya AR, et al. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. Nucleic Acids Research. 2022;51:D690–D699. doi:10.1093/nar/gkac920

77. Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, et al. ResFinder 4.0 for predictions of phenotypes from genotypes. Journal of Antimicrobial Chemotherapy. 2020;75:3491-3500. doi:10.1093/jac/dkaa345

78. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, et al. ARG-ANNOT, a New Bioinformatic Tool To Discover Antibiotic Resistance Genes in Bacterial Genomes. Antimicrob Agents Chemother. 2014;58:212-220. doi:10.1128/aac.01310-13

79. Bonin N, Doster E, Worley H, Pinnell LJ, Bravo JE, et al. MEGARes and AMR++, v3.0: an updated comprehensive database of antimicrobial resistance determinants and an improved software pipeline for classification using high-throughput sequencing. Nucleic Acids Research. 2022;51:D744–D752. doi:10.1093/nar/gkac1047

80. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep. 2021;11. doi:10.1038/s41598-021-91456-0

81. Blin K, Shaw S, Vader L, Szenei J, Reitz ZL, et al. antiSMASH 8.0: extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation. Nucleic Acids Research. 2025;53:W32–W38. doi:10.1093/nar/gkaf334

82. van Heel AJ, de Jong A, Song C, Viel JH, Kok J, et al. BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. Nucleic Acids Research. 2018;46:W278–W281. doi:10.1093/nar/gky383

83. Ehling-Schulz M, Fricker M, Grallert H, Rieck P, Wagner M, et al. Cereulide synthetase gene cluster from emetic Bacillus cereus: Structure and location on a mega virulence

plasmid related to Bacillus anthracis toxin plasmid pXO1. BMC Microbiol. 2006;6. doi:10.1186/1471-2180-6-20

84. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol. 2017;35:725-731. doi:10.1038/nbt.3893

85. Jiménez G, Urdiain M, Cifuentes A, López-López A, Blanch AR, et al. Description of Bacillus toyonensis sp. nov., a novel species of the Bacillus cereus group, and pairwise genome comparisons of the species of the group by means of ANI calculations. Systematic and Applied Microbiology. 2013;36:383-391. doi:10.1016/j.syapm.2013.04.008

86. Siguier P, Gourbeyre E, Chandler M. Bacterial insertion sequences: their genomic impact and diversity. FEMS Microbiol Rev. 2014;38:865-891. doi:10.1111/1574-6976.12067

87. Espino-Vázquez A, Gómez-Treviño A, Galán-Wong L, Pereyra-Alférez B. Isolation of Bacillus thuringiensis strains with cytotoxic activity against MOLT-4, a leukemia cell line. In: Microbes in Applied Research. Singapore: World Scientific; 2012. pp. 147-151. doi:10.1142/9789814405041_0029

Additional files

Additional file 1. File name: Additional_file_1_Figures_S1-S24.pdf. Format: PDF. Title of data: Supplementary figures S1–S24. Description: the twenty-four supplementary figures cited in the text, one per page in citation order (Fig. S8 is supplied as the two panels S8a and S8b on consecutive pages). Legends are listed below.

Supplementary figures

Fig. S1. Average nucleotide identity versus digital DNA–DNA hybridisation, all 190 genome pairs. Agreement between two independent species-delineation criteria across all 190 pairs: ANI against digital DNA–DNA hybridisation (GGDC formula 2). All 190 pairs resolved. The grey line is a degree-two polynomial fit of hybridisation on identity, over which the residual standard deviation is 1.112 percentage points; none of the 54 pairs involving a study strain falls outside three standard deviations of that fit. Dashed lines mark the equivalent species thresholds of the two criteria and the shaded area the region logically inconsistent between them. The two criteria are therefore mutually corroborating rather than merely co-reported.

Fig. S2. Closest relative of each genome by average nucleotide identity. Closest relative of each genome by average nucleotide identity, with the amber band marking the 95–96% range in which the criterion is least reliable and the dashed red line the 95% threshold. GM18 and IB84 exceed the threshold and are separated from their second-closest neighbour by 7.29 and 4.11 percentage points respectively. L4J2 is the exception: its two closest relatives, HD-789 and AM65-52, are themselves near-identical to it, so its margin to the second-closest genome is 0.01 points and no single nearest relative can be identified among the three.

Fig. S3. Maximum-likelihood phylogeny from twenty vertically inherited marker genes. Maximum-likelihood phylogeny from twenty vertically inherited marker genes (VBCG v1.3, RAxML, 100 bootstrap replicates), rooted on the same outgroup to permit comparison. Brackets mark the ANI species blocks, which cease to be clades in this tree: GM18 separates from the *B. toyonensis* type strain and AM65-52 leaves the L4J2 / HD-789 group. Three genome pairs were declared identical under this marker set, indicating loss of marker resolution rather than an alternative evolutionary history. This figure must not be presented as confirming Fig. 2. Branch lengths are not comparable in magnitude with Fig. 2; only topology is. Scale bar, 0.01 substitutions per site.

Fig. S4. Topological comparison of the core-genome and marker-gene phylogenies. Face-to-face comparison of the two phylogenies, drawn as cladograms because only topology is comparable between inferences on different character sets. Lines join the same genome in the two trees and their colour indicates the 95% ANI species block. The abundance of crossings reflects extensive topological discordance (Robinson–Foulds distance 26 of a possible 34; 4 of 17 internal nodes shared). The vertical ordering of taxa is chosen by the drawing algorithm to minimise crossings and has no biological meaning. The Robinson–Foulds distance counts unshared bipartitions without weighting by support, treating a 100% bootstrap node and a 23% node alike.

Fig. S5. Species clusters recovered as a function of the identity threshold applied. Single-linkage clustering of the twenty genomes at every ANI threshold between 90.0% and 99.5%. The orange series gives the number of genomes in the largest cluster and the blue series the total number of clusters recovered. The amber band marks the 95–96% range and the dashed line the conventional 95% threshold, at which a single cluster contains twelve of the twenty genomes, including the type strain of *B. cereus*, four *B. thuringiensis* reference strains and strain L4J2. GM18 and IB84 persist as two-genome clusters with their respective type strains throughout the range examined. Single linkage chains genomes through intermediates and is deliberately permissive: it is the appropriate choice for testing whether a threshold separates a set, and is not a classification proposal. The two steps in the blue series near 99% reflect the separation of near-clonal pairs and carry no taxonomic meaning.

Fig. S6. Parasporin-family gene families carried by each genome. Parasporin-like gene families carried by each genome under the structure-based App/Bmp/Mpp/Mtx/Sip/Spp/Tpp/Xpp nomenclature. Unlike the classical class, this one is nearly universal in the panel, with family Spp1 present in nineteen of the twenty genomes, so the distinguishing feature is the breadth of families and the confidence level reached. Sixty-three per cent of locus entries are distant matches below 45% identity and only eight per cent reach named-gene identity; any figure quoted for parasporin load must state the threshold applied. Selective cytotoxicity cannot be inferred from sequence similarity.

Fig. S7. Sequence-evidence strength behind each toxin locus assignment. Strength of the sequence evidence behind each detected locus, for the classical insecticidal class (a) and the structure-based class (b), broken down by the identity of the best BPPRC match into four tiers plus a category for loci detected only by hidden Markov model profile. The two classes differ markedly: 46% of classical loci reach named-gene identity against only 8% of structure-based loci, of which 63% are distant matches. The tiers are identity thresholds against a reference database, not measures of biological activity; a named locus may be pseudogenised. Horizontal axis scales differ between panels.

Fig. S8. Coding integrity of insecticidal toxin and parasporin-family loci in each genome. (a) Coding integrity of insecticidal toxin loci in each genome. (b) Coding integrity of parasporin-family loci in each genome. The 271 detected locus entries were classified by direct translation, checking length, internal stop codons and length ratio against the reference protein, and cross-referencing the result with two independent pseudogene call sets. Each bar corresponds to a genome and its total length to the number of loci detected; the text on the right gives the percentage of loci with an intact frame and the total on which it is computed. Rows follow the order of the core-genome phylogeny. All three loci of IB84 are intact; GM18 retains 12 of 18, with losses attributable to pseudogenisation that improved contiguity did not reverse; L4J2 retains 11 of 24 and includes independently confirmed pseudogenes, in the most fragmented assembly of the panel and the genome with by far the largest integrated prophage burden of the three study strains. Percentages for genomes with very few loci are fragile, which is why the denominator is always printed; the contig-edge category is an assembly artefact rather than a biological observation. The horizontal axis scales differ between panels.

Fig. S9. Replicon location of insecticidal and parasporin loci in each genome. Replicon location of toxin loci in each genome, separated by protein class. Loci were assigned to chromosome or plasmid by combining contig classification with RFPlasmid in the draft genomes, the declared sequence headers in the complete genomes, and an independent assignment with MOB-suite v3.1.9. The text on the right gives the percentage of loci on plasmids and the class total for that genome. In aggregate, 185 of 271 locus entries (68.3%) reside on plasmids. Replicon assignment is declared in the sequence record only for the 14 closed genomes; for the six drafts it is a prediction. That a gene lies on a plasmid does not imply it is mobilisable, which depends on the plasmid carrying conjugation or mobilisation systems (Fig. S14).

Fig. S10. Proportion of each gene family located on plasmids. Proportion of the loci of each gene family located on plasmids, aggregated across the twenty genomes, for the classical (a) and structure-based (b) classes. Point size is proportional to the number of loci and the dashed line marks the overall proportion of 68.3%. Families with fewer than three assigned loci are excluded, since small families give unstable proportions. The aggregation mixes complete genomes, where the replicon is declared, with drafts, where it is predicted. A family that is plasmid-borne in this panel need not be so throughout the group.

Fig. S11. Replicon location of each toxin family in each genome. Replicon location of each toxin family in each genome, showing locus counts and whether they reside on plasmid, chromosome or both. White cells indicate that the genome does not carry the family. A family may sit on different replicons in different genomes, which is what is expected of genes that move. Cells containing a single locus admit no intermediate category.

Fig. S12. Insertion sequences within 10 kb of each class of locus. For each class, the ratio of the observed number of loci with at least one insertion sequence within 10 kb to the number expected under random placement within the same replicon, shown with the observed and expected percentages. Elements were detected with ISEScan v1.7.3 and named by comparison with the ISfinder reference database; the analysis is restricted to the complete genomes, where contiguity permits reliable distance measurement, and loci close to one another are merged so that a single neighbourhood is not counted repeatedly. Toxin and parasporin-like loci lie closer to insertion sequences than expected, with 168 of them within 10 kb of an insertion sequence against 46% expected, whereas virulence factors, resistance genes and biosynthetic clusters do

not depart from expectation; bacteriocin genes also reach significance, though on a far smaller locus base. Association does not establish direction. The 10 kb window is one choice among three examined, which are nested and therefore not independent, and p-values are not corrected for multiple comparisons across the category, replicon and window combinations tested.

Fig. S13. Foreign mobile elements recorded in the CRISPR memory of each genome. Foreign mobile elements recorded in the CRISPR memory of each genome. Spacers were extracted with CRISPRCasTyper v1.9.0 and CRISPRDetect v2.4 and compared against the plasmids and prophages carried by the other genomes of the panel using CRISPRTarget and BLASTN, retaining only matches supported by both methods; matches against a genome's own elements are excluded. All twenty genomes are shown: the fifteen lacking a complete CRISPR-Cas system appear without a bar because they lack the machinery to record any memory at all, not because they encountered nothing. Of the seventy-one matches recovered, most correspond to plasmids of other genomes, documenting effective circulation of plasmid elements between lineages of the group. The analysis establishes neither the direction nor the date of any transfer, and a false-positive rate was bounded with a shuffled-spacer control.

Fig. S14. Plasmid content, largest plasmid and plasmid mobility class per genome. Plasmid content, largest plasmid and mobility class per genome. Plasmids were reconstructed and typed with MOB-suite v3.1.9, with independent typing by PlasmidFinder. The left panel gives total plasmid content in megabases, the centre panel the size of the largest individual plasmid in kilobases, and the right panel the number of plasmids of at least 10 kb classified as conjugative, mobilisable or non-mobilisable. Reconstruction from draft assemblies is incomplete, so counts are lower bounds, plasmid reconstruction from drafts being incomplete for at least 10 kb in the complete genomes. Mobility classification is predictive, based on the presence of the genetic determinants of conjugation rather than on a demonstration of transfer. The three panels use different units and should not be compared with one another.

Fig. S15. Antiviral defence systems and CRISPR-Cas content of each genome. Antiviral defence systems and CRISPR-Cas content per genome. The left panel gives the number of defence systems identified by reconciling DefenseFinder v3.0.0 and PADLOC v2.0.0, split by chromosomal or plasmid residence. The right panel gives the number of confident CRISPR repeats detected with CRISPRDetect v2.4 and marks in colour the genomes that additionally possess a complete CRISPR-Cas system typed with CRISPRCasTyper v1.9.0, with its subtype. Only five of the twenty genomes possess a complete system and all are subtype I-C; among the study strains only GM18, which also holds the largest number of plasmid-resident defence systems of the panel. Absence of a complete system is the majority state in this panel and must not be read as a deficiency of a particular strain. Detection counts are lower bounds, reflecting rapidly expanding databases.

Fig. S16. Virulence factor clusters detected in each genome. Virulence factor clusters detected in each genome (ABRicate v1.4.0 against VFDB, April 2026 release, after identity and coverage filtering). Columns are ordered by decreasing frequency in the panel. The repertoire is largely uniform and includes, in most genomes, the enterotoxigenic determinants characteristic of the *Bacillus cereus* group, also present in the reference strains used commercially as bioinsecticides. Presence of a virulence gene is not virulence: this figure must not be cited as a risk assessment, and detection depends on database coverage biased towards well-studied pathogens.

Fig. S17. Antimicrobial resistance gene families detected in each genome. Antimicrobial resistance gene families detected in each genome (ABRicate v1.4.0 against CARD, ResFinder, ARG-ANNOT, MEGARes and the NCBI catalogue, with calls reconciled by genomic coordinate to avoid counting one locus several times). The repertoire is dominated by intrinsic chromosomal β-lactamases of the *Bacillus cereus* group. Presence of a resistance gene is not phenotypic resistance, which can be established only by susceptibility testing, and this figure does not distinguish intrinsic from acquired determinants; the distinction relevant to biosafety is plasmid location.

Fig. S18. Biosynthetic gene clusters by class in each genome. Biosynthetic gene clusters per genome by chemical class (antiSMASH v8.0.4, full database). L4J2 has the highest number of regions in the panel, a trait shared with its closest relative AM65-52 and therefore characteristic of the clade rather than of the strain. A predicted region does not imply production of the metabolite, and region counts covary with assembly contiguity, so GM18's value should be read with that caveat.

Fig. S19. Bacteriocin genes by structural class in each genome. Bacteriocin genes per genome by structural class, from homology search against the class I, II and III reference protein panel of BAGEL4 and cross-checked against independently detected biosynthetic regions. The repertoire is modest and uniform across the panel. A sequence match does not imply bacteriocin production or antimicrobial activity, and the method detects only bacteriocins with sufficient similarity to the reference panel.

Fig. S20. Insertion sequence families in each genome. Insertion-sequence families present in each genome (ISEScan v1.7.3, named against ISfinder, complemented by a TnCentral search for transposons outside this class). Element density is substantially higher on plasmids than on chromosomes and is associated with pseudogenisation of toxin loci. In fragmented assemblies the count is inflated, because each contig edge generates truncated calls; GM18's value must not be compared directly with those of complete genomes. Detection does not distinguish active elements from degraded, immobile copies.

Fig. S21. Integrated prophages per genome and their replicon location. Integrated prophages per genome and their replicon location, from independent detection with geNomad v1.12.0 and PhiSpy v5.0.10 using the *Bacillus* model, with region completeness assessed by CheckV v1.1.1. L4J2 carries the highest prophage load of the panel, including several prophages integrated into plasmids. Prophage detection is predictive: it identifies regions with viral signature, not demonstrably inducible phages, and region boundaries are approximate.

Fig. S22. Pangenome compartments in each genome. Partition of the pangenome in each genome into core, accessory and strain-specific gene clusters (Panaroo v1.8.0 strict mode, complemented by the independent BPGA v1.3 partition). The size of the strain-specific compartment depends on the number of genomes in the panel and is not transferable to other studies; it also includes fragments and annotation artefacts in fragmented assemblies, which affects GM18 most.

Fig. S23. Strain-specific genes and their functional annotation. Strain-specific gene clusters per genome and the fraction with a functional assignment from eggNOG-mapper v2.1.15. In fragmented assemblies, genes split across contig boundaries inflate this count; L4J2, with 799 contigs, is the most affected genome of the panel. Lack of functional annotation does not imply novelty.

Fig. S24. Functional categories of the strain-specific genes. Functional categories of strain-specific genes, grouped into the four conventional divisions of the orthologous-group system. The denominator is the number of category assignments rather than the number of genes, since a gene with several category letters contributes to each. Poorly characterised categories dominate in every genome, as is usual for this compartment.

Additional file 2. File name: Additional_file_2_Tables_S1-S9.xlsx. Format: XLSX. Title of data: Supplementary tables S1–S9. Description: the nine supplementary tables, one worksheet per table, holding the per-pair, per-locus, per-family and per-strain values underlying the figures of the main text. The caption of each table is listed below.

Supplementary tables

Table S1. Pairwise genome relatedness for all 190 genome pairs. One row per pair of the twenty-genome panel, giving average nucleotide identity, digital DNA–DNA hybridization under formulas 1, 2 and 3, the dDDH value predicted from ANI by the degree-two polynomial fit, the standardised residual of that fit, and the audit fields used to verify the returned result set.

Table S2. Pesticidal protein gene content by strain and family. One row per strain and gene family (150 rows), giving the number of loci detected and the strongest sequence-evidence tier reached, for both the classical Cry/Cyt/Vip and the structure-based nomenclature panels.

Table S3. Genomic relatedness against toxin-repertoire overlap for the three study strains. One row for each study strain paired with every other genome of the panel (54 rows), giving average nucleotide identity, the Jaccard index over toxin family sets, and the smaller of the two family counts, which marks the pairs that are not interpreted as biological similarity.

Table S4. Deposited species name against genome-based assignment for the twenty genomes. One row per genome, giving the name under which the strain is deposited, the assignment returned by GTDB-Tk R232, and whether the two conflict.

Table S5. Replicon location of each toxin family in each strain. One row per nomenclature panel, family and strain (149 rows), giving the number of loci, how many received a replicon assignment, how many lie on plasmids, the resulting percentage, and whether the assignment is declared in the sequence record or predicted.

Table S6. Functional categories of the strain-specific genes. One row per strain and orthologous group category letter (311 rows), giving the collapsed functional division and the number of strain-specific gene clusters assigned to it.

Table S7. The twenty-genome panel. One row per genome, giving the deposited species name, the subspecies or serovar where applicable, the role the genome plays in the panel, assembly level, assembly accession, and the method by which it was retrieved from NCBI.

Table S8. Software inventory. One row per analysis step and tool (52 rows), giving the version used, its purpose, every non-default parameter, and the name and version of the reference database queried.

Table S9. Published qualitative phenotype beside the genomic calls for the three study strains. One row per strain, giving the insecticidal and other activities reported in the genome announcements, the named toxin classes detected here, and the fields recording which observations were excluded from interpretation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile2TablesS1S9.xlsx
- Additionalfile1FiguresS1S24.pdf