Food Chemistry: Molecular Sciences 12 (2026) 100338

Contents lists available at ScienceDirect

Food Chemistry: Molecular Sciences

journal homepage: www.sciencedirect.com/journal/food-chemistry-molecular-sciences

A WGS workflow for identifying genetically modified and foodborne-pathogenic *Bacillus* isolates

Maxime Godfroid [a], Alexander Van Uffelen [a,b,c], Marie-Alice Fraiture [a], Sigrid C.J. De Keersmaecker [a], Kevin Vanneste [a], Nancy H.C. Roosens [a], Bert Bogaerts [a,*]

[a] *Transversal activities in Applied Genomics, Sciensano, J. Wytmanstraat 14, 1050 Brussels, Belgium*
[b] *Department of Information Technology, Internet Technology and Data Science Lab (IDLab), Interuniversity Microelectronics Centre (IMEC), Ghent University, Ghent, Belgium*
[c] *Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium*

ARTICLE INFO

Keywords:
Surveillance
Genetically modified microorganism (GMM)
Bacillus cereus
Bacillus subtilis
Food safety
Bioinformatics workflow

ABSTRACT

Bacterial contamination of food and feed is an important public health issue that poses potential risks to consumers. Contamination can occur during industrial fermentation and production processes, where genetically modified micro-organisms (GMMs) and toxin-producing bacteria may be present. The *Bacillus* genus is particularly relevant in this context, as the *Bacillus subtilis* group is commonly used as GMM, while *Bacillus cereus* is often associated with foodborne outbreaks. Whole-genome sequencing (WGS) is a widely used method to detect and characterize foodborne pathogens, but comparatively little research has focused on its application to GMMs. Here, we present a WGS-based bioinformatics workflow for the characterization of *B. subtilis* group and *B. cereus* group isolates, which includes a novel approach for the detection of known GMMs based on detecting known transgenic elements and host strains. The workflow supports both short-read (Illumina) and long-read (Oxford Nanopore Technologies) sequencing data and performs common genomic assays such as quality checks or taxonomic identification. Additionally, isolates are screened for genes associated with antimicrobial resistance, virulence genes and mobile genetic elements. The workflow largely follows the recent EFSA guidelines for WGS-based characterization of micro-organisms in the food chain. We demonstrate that the workflow correctly identifies known genetically modified *B. subtilis* strains, while not mislabeling wild-type strains as GMM. Finally, using publicly available datasets, we show that the workflow accurately characterizes and identifies subspecies for *B. cereus*. This automated solution for detecting known GMMs and foodborne pathogens within the *Bacillus* genus can support regulatory compliance and contribute to ensure food safety.

1. Introduction

Bacterial contamination of food and feed products is a persistent potential threat to human and animal health. Enforcement agencies frequently report the presence of bacterial contaminants in fermentation products, including genetically modified (GM) *Bacillus subtilis* sensu lato (*s.l*). used as production strains (Deckers et al., 2020), and potentially pathogenic foodborne species such as *Bacillus cereus s.l.* (Ehling-Schulz et al., 2019). Historically, the identification of bacterial contamination in food and feed products has relied on targeted approaches, using either molecular or sequencing methods, which are mostly used on cultured isolates. In particular, one of the standard methods for enforcement laboratories to identify bacterial contamination is quantitative PCR (qPCR). In recent years, whole-genome sequencing (WGS) has become the method of choice for characterizing bacterial isolates, providing strain-level taxonomic identification and insights into antimicrobial resistance (AMR) or virulence factors encoding genes (Bogaerts, Nouws, et al., 2021; Kohl et al., 2018; Ortega-Sanz et al., 2023; Sherry et al., 2023). WGS has also been widely used to trace outbreaks and determine relatedness between samples (Bogaerts et al., 2023; Wang et al., 2023). In addition, the advent of third-generation long-read sequencing, such as Oxford Nanopore Technologies (ONT) sequencing, offers the possibility of reconstructing (nearly) closed genomes and complete plasmids with high accuracy, especially when combined with short-read data (Sanderson et al., 2024). The integration of WGS into the activities of many laboratories has led to guidelines for quality control and

* Corresponding author.
E-mail address: bert.bogaerts@sciensano.be (B. Bogaerts).

https://doi.org/10.1016/j.fochms.2025.100338
Received 13 June 2025; Received in revised form 5 December 2025; Accepted 7 December 2025
Available online 8 December 2025

validation of WGS-based analyses (Bogaerts et al., 2019; Kozyreva et al., 2017). However, guidelines for long-read sequencing remain scarce, given the rapid evolution of the technology.

In the European Union (EU) market, the use of genetically modified organisms (GMO), including genetically modified micro-organisms (GMM), in the food and feed chain is regulated under regulations (EC) No. 1829/2003 and 1830/2003, and is subject to prior authorization and risk assessment by the European Food Safety Agency (EFSA). In contrast, for fermentation products such as vitamins or food enzymes manufactured using GMMs as producer organisms, the presence of viable GMMs in the final product is currently not authorized on the EU market. However, the detection of these contaminations is inherently challenging for enforcement laboratories, as the genetic structure of the GMM constructs is often unknown (Fraiture, Bogaerts, et al., 2020). So far, two contaminant GMMs were isolated and fully described in the literature. The first one is a *Bacillus velezensis* strain modified to overproduce a protease, isolated from commercial food enzymes (Fraiture, Bogaerts, et al., 2020). This transgenic strain harbors a 6.7 kb episomal plasmid, derived from a pUB110 backbone, containing the *aadD* and *bleO* AMR genes associated with resistance to kanamycin and bleomycin, respectively, and a 2.3 kb protease-encoding gene. The second described GMM is a *B. subtilis* sensu stricto *(s.s.)* strain, modified to overproduce vitamin B_2 (also called riboflavin). The genetic modifications in this strain include: (1) deletion of the *ribDEA* genes from the chromosomal *ribDEAHT* operon; (2) disruption of the *recA* gene in the chromosome by insertion of the *cat* chloramphenicol resistance gene as a selection marker; (3) chromosomal integration of two GM plasmids, namely pGMsub01 and pGMsub02 (Berbers et al., 2020), which contain the AMR genes *aadD* (associated with aminoglycoside resistance) and *blaTEM-116* (associated with beta-lactam resistance) (Fraiture, Deforce, et al., 2020); and (4) a large 38.6 kb episomal plasmid that has been genetically engineered to overproduce vitamin B_2. This plasmid contains the complete *ribDEAHT* operon, as well as the AMR genes *blaTEM-116* and *erm(B)* (associated with erythromycin resistance) and *tet(L)* (associated with tetracycline resistance). It is worth noting that, a third *Bacillus* GMM was recently identified in the metagenomic sequencing data of commercial food enzymes (D'aes et al., 2022). This GMM was engineered to overproduce the amylase enzyme via a construct based on the pUB110 plasmid backbone. However, no isolate carrying the modified plasmid could be retrieved as it was found to be unculturable. Consequently, the host strain could not be fully characterized.

Furthermore, microbial fermentation processes can also introduce pathogenic species into products, as the fermentation conditions allow certain pathogens to thrive. Previous studies have identified viable *B. cereus* strains in fermentation products (Bogaerts et al., 2023). Consuming these products can lead to food poisoning, which can have severe symptoms. Therefore, monitoring the food chain is essential for detecting contamination and tracing the source of foodborne outbreaks (Bennett et al., 2013), thereby limiting the impact on public health. In particular, for *B. cereus s.l.*, several bioinformatics tools are available for species identification, typing, detection of antimicrobial resistance and virulence gene characterization based on WGS data (Carroll, Cheng, & Kovac, 2020; Díaz-Valerio et al., 2021; Liu et al., 2021; Shikov et al., 2020). One of the most commonly used tools is BTyper3, which performs taxonomic identification and complete isolate characterization of *B. cereus* strains (Carroll, Cheng, et al. 2020). However, because it was developed as a stand-alone tool, it does not perform other assays that are part of end-to-end automated workflows, such as pre-processing of the input data, quality control, plasmid detection, AMR gene detection, core-genome multi-locus sequence typing (cgMLST) and ribosomal MLST (rMLST) (Bogaerts et al., 2019; Jolley et al., 2012).

Compared to pathogen characterization and outbreak investigation, integration of WGS for routine GMM surveillance remains limited, despite its successful application in research settings (Barbau-Piednoir et al., 2015; Berbers et al., 2020; D'aes et al., 2021; Paracchini et al., 2017). Existing quality control and validation guidelines are primarily designed for pathogens and are not fully suited to detect GMMs or unauthorized bacterial contaminants in the food and feed chain. A key challenge in identifying transgenic strains, i.e., strains that have undergone intentional genetic modification, is that detecting the transgenic construct alone is insufficient (D'aes et al., 2021). Natural processes in bacteria, such as horizontal gene transfer (HGT) and recombination, can alter bacterial genomes organization, complicating identification (Arnold et al., 2022). Therefore, transgenic construct detection must be coupled with strain identification to reliably link a GMM to its host background. In 2024, EFSA published guidelines to help conducting risk assessment studies on the safety of products and enforce legislation on the EU market by harmonizing the WGS analysis across laboratories. According to this document, the main requirements for WGS-based GMM analysis are: (1) exact strain identification; (2) characterization of genetic modifications (if any) and; (3) characterization of genes of concern, such as genes encoding virulence factors or genes associated with AMR (European Food Safety Authority (EFSA), 2024).

In this study, we present a WGS-based bioinformatics workflow for the identification and characterization of GM *B. subtilis s.l.* and *B. cereus s.l.* strains commonly isolated as contaminant from food and feed products, supporting both short-read (Illumina) and long-read (ONT) sequencing. This workflow can perform the identification of known GM *B. subtilis s.l.* using a novel approach that combines the detection of transgenic constructs with strain identification. In addition, the workflow is able to identify and fully characterize *B. cereus s.l.* to assess their pathogenic potential. By extensively validating the performance of the workflow, we demonstrate that it is suitable for use within quality systems under which many enforcement laboratories operate. Our work promotes the standardized use of WGS as a tool for public health surveillance, thereby improving the safety of the food chain.

2. Materials and methods

2.1. Bioinformatics workflow

The bioinformatics workflow starts from raw FASTQ input data, which can be generated using either Illumina or ONT sequencing. When both Illumina and ONT data are provided, the workflow uses a hybrid assembly approach. The targeted species (i.e., *Bacillus subtilis s.l.* or *Bacillus cereus s.l.*) must be specified. The workflow starts with a set of common pre-processing and quality control (QC) steps, followed by *de novo* assembly and a set of species-specific assays (Fig. 1). Finally, the workflow generates an HTML output report containing the most relevant information. An additional tabular summary file is generated that contains the same information as the HTML report, in a format that can easily be processed by command line tools or spreadsheet software.

2.1.1. Implementation and availability

The bioinformatics workflow was implemented in Python v3.10 and uses Snakemake v7.32.4 for parallelization (Mölder et al., 2021). The workflow was developed on Ubuntu v22.04. It is available for academic and non-profit use through the Galaxy instance of our institute at https://galaxy.sciensano.be (registration required) (Bogaerts et al., 2025).

2.1.2. Data preprocessing, quality filtering and de novo assembly

The workflow starts with an optional human read removal step using the NCBI human read removal tool v2.2.1 with default options for both Illumina and ONT data (NCBI, 2023). This step is recommended to ensure that the results of subsequent analyses are not affected by reads of human origin, which may be present due to contamination or the source of the sample. Although the latter is unlikely for isolates obtained from food or feed. For hybrid input data, both long and short reads are processed separately. This optional step is recommended to ensure that the results of subsequent analyses are not affected by reads of human origin, which may be present due to contamination or the source of the

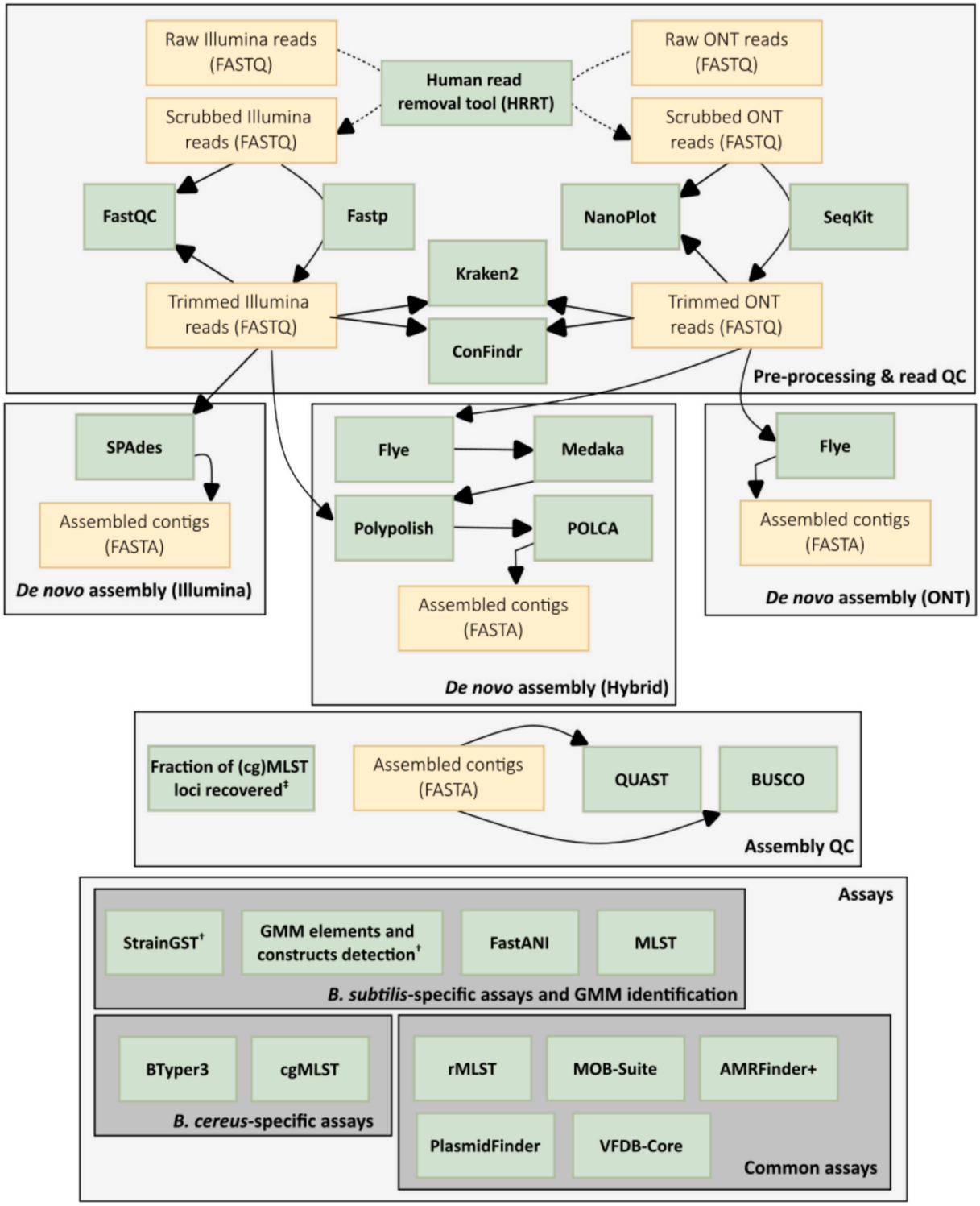

Fig. 1. Overview of the bioinformatics workflow. Input reads are first optionally (represented by a dotted line) scrubbed of human reads, then trimmed and assembled. Illumina and ONT reads are assembled using SPAdes and Flye, respectively. If both are provided, a hybrid is approach is used, where the Flye assembly is first polished with the long reads using Medaka, followed by short read polishing with Polypolish and POLCA. The quality of the input dataset is assessed by a set of quality checks (Table S1). The additional assays, some of which are species-specific, are either run on the resulting assembly or directly on the reads (marked with a †). The input for the assay marked by ‡ is retrieved from the (cg)MLST assay. The PubMLST cgMLST scheme is used for *B. cereus*, while the conventional MLST scheme is used for *B. subtilis*, since a cgMLST scheme is not available for this species. Finally, the workflow exports a final HTML report summarizing the main results, as well as a tabular output file that can easily be further processed. A simplified workflow summarizing the main steps is provided in Fig. S4. Abbreviations: genetically modified micro-organism (GMM), quality control (QC).

sample.

2.1.2.1. Illumina data. The paired-end Illumina reads are trimmed using fastp v0.23.4 (Chen, 2023) with the following options: '–detect_adapter_for_pe', '–cut_front', '–cut_front_window_size' set to 1, '–cut_front_mean_quality' set to 10, '–cut_tail', '–cut_tail_window_size' set to 1, '–cut_tail_mean_quality' set to 10, '–cut_right', '–cut_right_window_size' set to 4, '–cut_right_mean_quality' set to 20, and '–length_required' set to 40. The quality of the reads is then evaluated using FastQC v0.11.7 (Simon, 2010). Additionally, ConFindr v0.8.2 is used to screen for intra-species and cross-species contamination based on the rMLST database ('–rmlst') and other options left at default values (Low et al., 2019). Finally, the processed reads are assembled de novo using SPAdes v3.15.5 with the '–isolate' option enabled and '–cov-cutoff' set to 10 (Prjibelski et al., 2020). Contigs smaller than 1 kb are removed using the 'seq' function from seqtk v1.4 (available at https://github.com/lh3/seqtk).

2.1.2.2. ONT data. The ONT input reads are trimmed using SeqKit v2.3.1 with the '–min-len' option set to 1000 and the '–min-qual' set to 10 (Shen et al., 2016; Shen et al., 2024). The quality of the reads is then evaluated using NanoPlot v1.41.6 (De Coster & Rademakers, 2023) and SeqKit v2.3.1 (Shen et al., 2016; Shen et al., 2024). Then, ConFindr v0.8.2 is executed with options '–data-type' set to 'Nanopore' and '–quality_cutoff' set to 12 (Low et al., 2019). Finally, the processed reads are de novo assembled using Flye v2.9.4, providing the reads using '–nano-corr' option and other options left at default values (Kolmogorov et al., 2019).

2.1.2.3. Hybrid data. For hybrid data (i.e., Illumina and ONT data processed together), a long read-first assembly approach is used (Wick et al., 2023). The assembly steps for the ONT data are first performed as described above. Then, Medaka v1.11.3 (Nanoporetech, 2024) is used to polish the resulting assembly with the trimmed ONT reads using default parameters. Finally, Polypolish v0.6.0 (Wick & Holt, 2022) and POLCA v4.1.0 (Zimin & Salzberg, 2020) are used sequentially, both with default parameters, to polish the resulting assembly using the trimmed Illumina reads.

2.1.2.4. Quality checks. In addition to the QC checks on the reads, the workflow screens for contaminants at the genus level (i.e., reads not assigned to the *Bacillus* genus) using Kraken2 v2.1.1 on the trimmed reads (Wood et al., 2019). The Kraken2 database contains all NCBI RefSeq (O'Leary et al., 2016) 'Genome' entries (database accessed February 24th, 2024) annotated as 'complete genome' with accession prefixes NC, NW, AC, NG, NT, NS and NZ of the following taxonomic groups: archaea, bacteria, fungi, protozoa and viruses. The database also contains the human genome (accession: GCF_000001405.40). Furthermore, the quality and completeness of the assembly is assessed by: (1) QUAST v5.2.0 (Mikheenko et al., 2018); (2) the fraction of typing loci recovered (core-genome multi-locus sequence typing ((cg)MLST) for *B. cereus*, MLST for *B. subtilis*) and; (3) estimating the completeness of the genome by screening for near-universal single-copy orthologs using BUSCO v5.5.0 (Manni et al., 2021). For QC checks that fail, message(s) are included in the HTML and summary outputs to inform the users. The workflow continues even if a QC check fails, but the user should be cautious in interpreting the results. An overview of the quality checks and their warning and failure thresholds is provided in Table S1.

2.1.3. Common assays

Several assays are executed for both species (i.e., *B. cereus s.l.* and *B. subtilis s.l.*). First, AMRFinder+ v3.11.26 (Feldgarden et al., 2021) is used to screen for genes associated with AMR using the assembly as input. In addition, the assemblies are screened for virulence genes and plasmid replicons using the Virulence Factor core database (VFDB) (Liu et al., 2022) and the PlasmidFinder Gram-positive database (Carattoli et al., 2014), respectively. The detection is performed using a BLAST+-based approach (v2.14.0) described previously (Bogaerts, Nouws, et al., 2021; Camacho et al., 2009). Then, MOB-suite v3.1.8 (Robertson et al., 2020; Robertson & Nash, 2018) is used with default parameters to screen for putative mobile genetic elements (MGEs). MOB-suite groups contigs that are assigned to the same putative MGE, which is used to the predict the genomic context of the detected AMR and virulence genes by cross-checking the corresponding contigs (i.e., whether they are located on the chromosome or on a MGE as determined by MOB-suite). Finally, sequence types and subspecies are determined using a sequence typing approach described previously (Bogaerts, Nouws, et al., 2021), using the MLST and rMLST profiles obtained from PubMLST (Jolley et al., 2012). The AMRFinder+, PlasmidFinder and typing databases are updated on a weekly basis to ensure up-to-date results. The workflow reports include the dates of the last database update.

2.1.4. Species-specific assays (B. cereus)

For *B. cereus s.l.*, the workflow includes characterization by BTyper3 v3.4.0, which performs MLST, phylogenetic group assignment using the pantoate-β-alanine ligase (*panC)* gene sequence typing, (sub)-species identification by average nucleotide identity (ANI) and screening for genes encoding virulence factors using an accompanying database (Carroll, Cheng, & Kovac, 2020; Carroll, Wiedmann, & Kovac, 2020). The PubMLST cgMLST scheme for *B. cereus* is used for additional sequence typing (Jolley et al., 2018; Tourasse et al., 2023).

2.1.5. Species-specific assays (B. subtilis)

2.1.5.1. (Sub-)species identification and typing. For *B. subtilis*, the workflow performs (sub)species identification by ANI using FastANI v1.33 with default parameters on the assembled contigs (Jain et al., 2018). A custom ANI database was built from 1624 *Bacillus* genomes from the RefSeq database (O'Leary et al., 2016) (taxid: 55087, assembly level of at least chromosome, accessed on May 9th, 2024). The ten best matching genomes from the reference genome database are reported in the HTML report and summary output, along with the associated taxonomic metadata. Since no cgMLST scheme is currently available for *B. subtilis*, this assay is not included for this species, but the rMLST and MLST schemes are used.

2.1.5.2. GMM characterization. The workflow performs the identification of known fully characterized GM strains using a two-step approach. First, the reads are screened against a databases of known GM transgenic elements, such as selection markers, transgenic junctions and complete GMM constructs. An overview of the entries in the database is provided in Table 1, with more detailed descriptions in Table S2. Second, strain identification is performed by comparison of the reads with a database of reference strains. If the combination of the detected GM transgenic element(s) and the strain matches a known transgenic strain in the database, the workflow will report a match to the identified GMM. Note that this is a targeted approach and transgenic strains that are not in the database (e.g., another strain with the same transgenic construct) will hence not be flagged as known GMMs. Nevertheless, the workflow will report the hit(s) on the transgenic construct or the match with the transgenic strain, even if the combination does not match a database entry. Currently, the database contains the pUB110-protease (carrying the AMR-associated genes *aadD* and *bleoR*) and the pGM-rib (carrying the AMR-associated gene *cmR1*) constructs, the only GM constructs that are fully characterized at the genomic level so far, but the database will be expanded as data for other transgenic strains becomes available. The GMM engineered to overproduce amylase is not included in the database as the host strain could not be fully characterized in the previous study (D'aes et al., 2022). The workflow includes an additional screening for ~500 bp sequences corresponding to unnatural junctions from fully characterized GMMs (i.e., between the wild-type (*wt)* strain genome and

Table 1
Summary of the GMM transgenic elements database entries.

Entry ID	Description	GenBank Accession (if available)	Length (bp)	GMM
pUB110	Shuttle vector used in the protease GMM construct	M19465.1	4548	pUB110-protease
pHY300PLK	Shuttle vector used in GMM	n/a	4870	n/a
*kanR*1-*aadD*	Kanamycin resistance gene	n/a	762	pUB110-protease
*kanR*2-Kanamycin phosphotransferase	Kanamycin phosphotransferase resistance gene	n/a	813	pGM-rib, pUB110-protease, pUB110-amylase
*cmR*1-Chloramphenicol	Chloramphenicol resistance gene	n/a	651	pGM-rib
*cmR*2-Chloramphenicol	Chloramphenicol resistance gene	n/a	660	pUB110-amylase
bleoR-Bleomycin	Bleomycin resistance gene	n/a	399	pUB110-protease, pUB110-amylase
*eryR*1-ErythromycinB	Erythromycin B resistance gene	n/a	764	pGM-rib
*eryR*2-ErythromycinC	Erythromycin C resistance gene	n/a	735	pGM-rib
*tetR*1-TetracyclineL	Tetracycline L resistance gene	n/a	1377	pGM-rib
*tetR*2-TetracyclineC	Tetracycline C resistance gene	n/a	1191	pGM-rib
*ampR*1-Betalactamase	Beta-lactamase resistance gene	n/a	861	Used in GM bacteria (Fraiture, Deckers, et al., 2020)
*ampR*2-Betalactamase	Beta-lactamase resistance gene 2	n/a	861	Used in GM bacteria (Fraiture, Deckers, et al., 2020)

The first column lists the identifiers. The second column provides the corresponding descriptions, the third column provides the GenBank accession number (if available), the fourth column displays the length of the entry in base pairs (bp) and the fifth column displays the GMM associated with each element. If an element has been observed in a contaminated sample but is not associated with a particular GMM construct, this is indicated. Abbreviations: not available (n/a).

the GM construct) (Table 2), as well as junctions from the incompletely characterized GMM for which no isolate data is available (D'aes et al., 2022). These hits are reported in the output report, but are not considered for the GM identification assay, which is based solely on the identification of the complete GM plasmidic construct and the identification of the host strain.

For the fully characterized pUB110-protease and pGM-rib constructs, the Rep1 and Inc18 replicons were identified using the PlasmidFinder database, respectively. The pUB110-protease construct was assigned to primary cluster group AC670 by MOB-suite. In contrast, the pGM-rib contig was not classified as plasmidic by MOB-suite, despite the detection of the Inc18 replicon.

2.1.5.3. Detection of GM transgenic elements. The detection of the transgenic elements is performed using KMA v1.4.12a on the trimmed reads, as described previously (Bogaerts, Nouws, et al., 2021; Clausen et al., 2018). Transgenic elements are considered present if they are covered for over 80 % of the sequence with over 80 % nucleotide identity. For ONT reads, the additional parameter '-bcNano' is enabled and the parameter '-bc' is set to 0.7. For hybrid datasets, KMA is used on the Illumina and ONT reads separately, and a sequence is considered present if it is detected in both sets of reads.

2.1.5.4. Strain identification. StrainGST v1.3.9, with default options, is used for strain identification. StrainGST uses a k-mer-based approach to compare the input dataset to a genome database (van Dijk et al., 2022). Although the tool was originally developed for Illumina data and does not explicitly support ONT data, it was able to analyze the ONT datasets and obtain identical results to those obtained with Illumina data (see section 2.2). The output of StrainGST includes: (1) the breadth of coverage (i.e., the k-mer fraction of the sample present in the reference); (2) the evenness of coverage along the reference genome (ranging from 0 to 1, where 1 corresponds to the reference genome being evenly covered); (3) the estimated relative abundance of the strain and (4) the score, ranging from 0 to 1, which represents a rank of the references in the database. A high score indicates a high confidence in the inferred strain. Note that for isolate data, the expected relative abundance is 100 %.

The database was constructed by collecting 1624 Bacilli genomes from RefSeq with assembly levels 'chromosome' or 'genome', using the StrainGST 'ncbi-genome-download' and 'prepare_strainge_db' modules (executed on March 9th, 2024 following the authors guidelines). Note that this selection of genomes is identical to the database used for species identification by ANI (Section 2.1.4). The genomes were then clustered based on similarity measures. In short, genomes with more than 99 % of k-mers present in another genome were discarded, and the remaining genomes were clustered together if their Jaccard similarity index was greater than 0.9. The genome with the smallest average distance to the other cluster members was used as the representative genome for each cluster. Finally, the cluster representatives were k-merized using the StrainGST 'kmerize' module (with default parameters). The genome of the transgenic *B. velezensis* strain containing the pUB110-protease transgenic construct and the genome of the *B. subtilis* containing the pGM-rib transgenic construct were additionally included in the database. The representative genome of the pUB110-protease cluster is denoted as "Baci_velezensis_10075", and the representative genome of the pGM-rib cluster is denoted as "Baci_subtilis_LBUM979".

In hybrid mode, the assay is run on both the Illumina and ONT reads, and a warning is issued if there is a discordance between both. Furthermore, if the assay detects a known GMM strain in at least one of the read sets, it will report a match to that strain.

2.1.6. Concordance with EFSA regulations for data analysis

The workflow has been developed largely in accordance with the EFSA guidelines for reporting whole genome sequence analysis of microorganisms intentionally used in the food chain (European Food Safety Authority (EFSA), 2024). First, the BUSCO completeness and total assembly length deviation checks included in the EFSA guidelines are enforced in our workflow (Table S1). Furthermore, the workflow includes several additional quality metrics and QC checks that are not listed in the EFSA guidelines (Table S1). Second, in accordance with the EFSA guidelines, strain identification by ANI is included as an assay in the workflow. Third, the EFSA recommends that, in order to report genes and genetic elements of concern, these should be covered for at least 70 % of their length with at least 80 % nucleotide identity. Here, we applied slightly more stringent thresholds, which are commonly used, e.g., for AMR gene detection (Bogaerts et al., 2019; Bogaerts, Nouws, et al., 2021). We also chose not to follow the recommendation to use at least two separate AMR databases, in order to reduce redundancy in the output and simplify interpretation. Finally, the detection of genetic modification differs from the EFSA guidelines in that it is target-based (i.e., linked to a predefined database), whereas the guidelines recommend reporting all genetic modifications from the *wt* strain, which would be very difficult to fully automate.

Table 2
Overview of the transgenic junctions in the database.

Entry ID	Description	Length (bp)	GMM
pUB110-protease	Complete plasmid sequence of the pUB110-protease construct	6756	pUB110-protease
pUB110-amylase	Complete plasmid sequence of the GMM-amylase construct	6814	pUB110-amylase
pGMrib	Complete plasmid sequence of the pGM-rib construct	38,647	pGM-rib
GMMprotease1_L	Left junction of pUB110 with protease on pUB110-protease, spanning 250 bp downstream and upstream of the junction.	500	pUB110-protease (D'aes et al., 2021)
GMMprotease1_R	Right junction of protease with pUB110 on pUB110-protease, spanning 250 bp downstream and upstream of the junction.	500	pUB110-protease (D'aes et al., 2021)
GMMprotease2_L	Left junction of pUB110 with *nprE*	461	GMM protease 2 (Fraiture et al., 2021)
GMMprotease2_R	junction of *nprE* with pUB110	454	GMM protease 2 (Fraiture et al., 2021)
GMMamylase1_L	Left junction of pUB110 with amylase on pUB110-amylase, spanning 250 bp downstream and upstream of the junction.	500	GMM amylase 1 (D'aes et al., 2022)
GMMamylase1_R	Right junction of amylase with pUB110 on pUB110-amylase, spanning 250 bp downstream and upstream of the junction.	500	GMM amylase 1 (D'aes et al., 2022)
GMMamylase2_L	Junction of *catA* with *amyS* on chromosome *B. licheniformis* host	500	GMM amylase 2 (Fraiture et al., 2024)
GMMamylase2_R	Junction of *amyS* with *catA* on chromosome *B. licheniformis* host	500	GMM amylase 2 (Fraiture et al., 2024)
GMMvitb2_558	Junction of *catA* with *recA* on *B. subtilis* host	500	pGM-rib (Paracchini et al., 2017)
GMMvitb2_690	Junction of pUB110 to pUC19 on *B. subtilis* host	500	pGM-rib (Paracchini et al., 2017)
GMMvitb2_691	Junction of *B. amyloliquefaciens* rib operon to pUC19 on *B. subtilis* host	500	pGM-rib (Paracchini et al., 2017)
GMMvitb2_804	Junction of deleted *B. amyloliquefaciens* rib-operon on *B. subtilis* host	500	pGM-rib (Paracchini et al., 2017)
GMMvitb2_693	Junction of 30-ribfragment to pUC19 on pGM-rib	500	pGM-rib (Paracchini et al., 2017)
GMMvitb2_694	Junction of riboperon-fragment to pSM19035 on pGM-rib	500	pGM-rib (Paracchini et al., 2017)

The transgenic junctions displayed in this table are used to identify a specific GMM. The first column lists the identifiers, the second column provides the corresponding descriptions (referenced in the associated papers), the third column displays the length of the entry in base pairs (bp) and the fourth column displays the GMM associated with the junction. Abbreviations: not available (n/a).

2.2. Performance evaluation

The workflow performance was evaluated using datasets sequenced in the context of this study, complemented with public data (Table S12). The complete dataset is described in the following section. For *B. subtilis*, the GMM detection assay was evaluated based on: (1) identification of the complete GMM pUB110-protease construct in GMM isolates and no detection of GMM constructs in *wt* strains; (2) identification of the correct GMM strain cluster of (i.e., the cluster represented by the genome denoted as "Baci_velezensis_10075" by StrainGST). In addition, the detection of the AMR-associated genes *bleO* and *aadD*, both part of the GMM pUB110-protease construct, was evaluated. For the *wt B. velezensis* samples, the absence of GMM and the associated AMR genes was evaluated.

As an independent test case, the WGS data for the pGM-rib GMM *B. subtilis* strain was analyzed, which contains a large 38.6 kb plasmid genetically modified to overproduce riboflavin (i.e., the pGM-rib GMM). We evaluated whether the GMM detection assay correctly reported the full pGM-rib construct, as well as the strain cluster represented by the "Baci_subtilis_LBUM979" genome.

For *B. cereus*, the assays that were evaluated were typing, sub-species identification, and the detection of toxin-encoding genes, more specifically *Bt* toxin-encoding genes (referred to as "*Bt* genes"), as described in section 2.2.1.2.

2.2.1. Datasets

2.2.1.1. B. subtilis s.l.. The performance was evaluated using WGS data from independent strains (i.e., separate cultures) of the previously described GMM pUB110-protease, i.e., a genetically modified *B. velezensis* strain, containing a GM insert for overexpression of a protease. The isolates were obtained from four different products, in which the same GM strain was identified, as described previously (D'aes et al., 2021). In total, ten WGS datasets were used (three isolates with two biological replicates, one isolate with four biological replicates). The strain was assigned to ST140 using the *B. subtilis* MLST scheme from PubMLST. These GMM constructs were originally characterized using a combination of targeted qPCR and hybrid sequencing (D'aes et al., 2021). As negative controls, six different *wt B. velezensis* strains were analyzed with the workflow to verify that they were not identified as GMMs. Genomic DNA of the *wt B. velezensis* strains were obtained from the Belgian Coordinated Collection of Microorganisms (https://bccm.belspo.be/) and were sequenced using the Illumina MiSeq (see below). One isolate of the previously described GMM pGM-rib, i.e., a *B. subtilis s.l.* strain containing a GM insert for the overexpression of vitamin B2, was used as an independent test case. The sequencing data for this strain is publicly available (Table S12).

2.2.1.2. Bacillus cereus s.l.. The performance for the *B. cereus* assays was evaluated using a collection of 57 *B. cereus* sensu stricto (*s.s.*) biovar *Thuringiensis* samples, for which hybrid sequencing data is available (Chung et al., 2024). These samples were originally part of the authors' lab culture collection and were collected mostly from environmental and food matrices (Table S3). The samples were phenotypically tested for toxin production and then screened for the presence of *Bt* genes using several bioinformatics methods. Note that this study only reported the presence or absence of *Bt* genes, without specifying particular genes. We evaluated the performance of the workflow on (1) the concordance between the species assignment of the original study and our workflow and (2) the accurate detection of *Bt* genes in the *B. cereus s.s.* samples compared to the original study. The performance was evaluated separately for the ONT, Illumina, and hybrid data. This dataset was supplemented with nine in-house *B. cereus s.l.* samples obtained from alcoholic beverage processing to evaluate the performance of the (sub) species identification. The isolates were selected after overnight plating of the enzymatic solution, after which the selected isolates were grown overnight and DNA was extracted. The isolates were then sequenced using the Illumina MiSeq (see below).

Since the nine in-house validation datasets lacked species labels, an independent method using a clustering approach was used to generate the truth set to validate the performance of the species identification assay of the workflow. All available genomes from BTyperDB were downloaded (accessed on August 2nd, 2024), totaling 5976 genomes (Ramnath et al., 2023). Genomes lacking valid taxonomic labels were excluded, resulting in 34 removed genomes, labeled as 'Species unknown' (n = 21), 'B. UnknownSpecies13' (n = 10), 'B.

UnknownSpecies15' (n = 2), and 'B. UnknownSpecies18' (n = 1). Note that these labels are placeholders used by BTyperDB. The remaining 5942 genomes, along with the samples included in this study, were subjected to a pairwise ANI calculation with FastANI v1.33 using default settings (Jain et al., 2018). These ANI values were used to hierarchically cluster the BtyperDB genomes and the new *B. cereus* assemblies. Clustering was performed in Python 3.10 using the 'clustermap' function from Seaborn v0.13 with default settings (Waskom, 2021). Species clusters, excluding the 'unknown' genomes, were demarcated by minimizing the distance between clusters, while ensuring that each species was confined to a single cluster. In other words, the dendrogram was divided into the maximum number of clusters possible, ensuring that no species spanned more than one cluster. Species labels for the analyzed samples were then assigned based on the exact location in this clustering.

2.2.2. DNA sequencing

The GMM *B. velezensis* isolates and the GMM *B. subtilis* isolate were previously sequenced using Illumina and ONT (R9 technology) sequencing, and the publicly available data was used for this study (Berbers et al., 2020; D'aes et al., 2021). The accession numbers for these datasets are provided in Table S12. For the *wt B. velezensis* samples

Table 3
WGS workflow results for the genomes used in the *B. subtilis* performance evaluation.

Isolate	Type	Sequencing technology	AMR genes	StrainGST		GMM			Hybrid	
				Strain	Score	Construct	Id. (%)	Cov. (%)	Id. (%)	Cov. (%)
cob91	GMM	Illumina	*aadD1, bleO, satA**	Baci_velezensis_10075	0.924	pUB110-protease	99.99	99.99	100	100
cob91	GMM	ONT	*aadD1, bleO, satA**	Baci_velezensis_10075	0.905	pUB110-protease	100	100		
cob92	GMM	Illumina	*aadD1, bleO, satA**	Baci_velezensis_10075	0.920	pUB110-protease	99.99	99.99	100	100
cob92	GMM	ONT	*aadD1, bleO, satA**	Baci_velezensis_10075	0.904	pUB110-protease	100	100		
pure1	GMM	Illumina	*aadD1, bleO, satA**	Baci_velezensis_10075	0.921	pUB110-protease	99.99	99.99	100	100
pure1	GMM	ONT	*aadD1, bleO, satA**	Baci_velezensis_10075	0.905	pUB110-protease	100	100		
pure2	GMM	Illumina	*aadD1, bleO, satA**	Baci_velezensis_10075	0.924	pUB110-protease	100	100	100	100
pure2	GMM	ONT	*aadD1, bleO, satA**	Baci_velezensis_10075	0.898	pUB110-protease	100	100		
crystal1	GMM	Illumina	*aadD1, bleO, satA**	Baci_velezensis_10075	0.923	pUB110-protease	100	100	100	100
crystal1	GMM	ONT	*aadD1, bleO, satA**	Baci_velezensis_10075	0.903	pUB110-protease	100	100		
crystal2	GMM	Illumina	*aadD1, bleO, satA**	Baci_velezensis_10075	0.924	pUB110-protease	100	100	100	100
crystal2	GMM	ONT	*aadD1, bleO, satA**	Baci_velezensis_10075	0.909	pUB110-protease	100	100		
pilsner11	GMM	Illumina	*aadD1, bleO, satA**	Baci_velezensis_10075	0.925	pUB110-protease	100	100	100	100
pilsner11	GMM	ONT	*aadD1, bleO, satA**	Baci_velezensis_10075	0.901	pUB110-protease	100	100		
pilsner12	GMM	Illumina	*aadD1, bleO, satA**	Baci_velezensis_10075	0.925	pUB110-protease	100	100	100	100
pilsner12	GMM	ONT	*aadD1, bleO, satA**	Baci_velezensis_10075	0.897	pUB110-protease	100	100		
pilsner21	GMM	Illumina	*aadD1, bleO, satA**	Baci_velezensis_10075	0.912	pUB110-protease	99.99	99.99	100	100
pilsner21	GMM	ONT	*aadD1, bleO, satA**	Baci_velezensis_10075	0.906	pUB110-protease	100	100		
pilsner22	GMM	Illumina	*aadD1, bleO, satA**	Baci_velezensis_10075	0.921	pUB110-protease	100	100	100	100
pilsner22	GMM	ONT	*aadD1, bleO, satA**	Baci_velezensis_10075	0.907	pUB110-protease	100	100		
pGM-rib	GMM	Illumina	*aadD1, tet(L), blaTEM-116, erm (B), catA, aadK, vmlR, mphK, bleO**	Baci_subtilis_LBUM979	0.959	pGM-rib	91.9	91.9	99.88	99.96
pGM-rib	GMM	ONT	*aadD1 aadK*, mphK*, catA*, blaTEM*, erm(B)*, tet(L)*, blaTEM**	Baci_subtilis_LBUM979	0.912	pGM-rib	99.99	99.99		
LMG12384	*wt*	Illumina	*satA*, clbA**	Baci_amyloliquefaciens_PM415	0.668	n/a	n/a	n/a	n/a	n/a
LMG17599	*wt*	Illumina	*satA**	Baci_velezensis_HC-8	0.788	n/a	n/a	n/a	n/a	n/a
LMG22478	*wt*	Illumina	*satA**	Baci_velezensis_SRCM123815	0.721	n/a	n/a	n/a	n/a	n/a
LMG23203	*wt*	Illumina	*satA**	Baci_velezensis_CACC_316	0.917	n/a	n/a	n/a	n/a	n/a
LMG26770	*wt*	Illumina	*clbA*	Baci_amyloliquefaciens_GL18	0.994	n/a	n/a	n/a	n/a	n/a
LMG27586	*wt*	Illumina	*satA*, clbA**	Baci_velezensis_AP3	0.993	n/a	n/a	n/a	n/a	n/a

For each isolate (first column), the columns show in order: the type of sample (either GMM or wt), the sequencing technology, the AMR genes detected by AMRFinder+ (imperfect hits are marked with a star), the closest strain identified by StrainGST (the naming follows the StrainGST nomenclature), the StrainGST score, the GMM construct detected, the sequence identity to the known GMM construct and the percent of the known GMM construct that is covered. The last two columns show the sequence identity to the known GMM construct and the percent of the known GMM construct that is covered in the hybrid assemblies. Abbreviations: GMM (genetically modified microorganism), wt (wild-type), n/a (not applicable).

and the *B. cereus s.l.* datasets (excluding the *B. cereus s.s.* samples obtained from another study), DNA from bacterial isolates was extracted using Quick-DNA™ HMW MagBead Kit (Zymo Research, Irvine, CA, USA) according to the manufacturer's instructions and then sequenced on the Illumina platform, following previously described protocol (D'aes et al., 2021). Short-read DNA libraries were prepared using the Nextera XT DNA library preparation kit (Illumina, San Diego, CA, USA) according to manufacturer's instructions. Sequencing was carried out on an Illumina MiSeq system with the V3 chemistry, obtaining 250 bp paired-end reads, and aiming for a theoretical coverage of 60× per sample, based on an average *Bacillus* genome size of 4.2 Mbp. The Illumina and ONT (R9) data for the *Bacillus cereus s.s.* samples from the study by Chung et al. were retrieved from SRA (Chung et al., 2024).

3. Results

3.1. Performance evaluation – B. subtilis

3.1.1. GMM B. velezensis strains

3.1.1.1. QC checks. All GMM datasets passed the QC checks included in the bioinformatics workflow. Of note, the number of contigs for the ONT and hybrid datasets ranged from 9 to 15, and from 27 to 39 contigs for the Illumina datasets. In each sample, the chromosome was fragmented into multiple contigs, indicating that complete circular chromosome sequences could not be obtained for any of the datasets (Table S4).

3.1.1.2. GMM construct and elements detection. For all GMM hybrid and ONT samples, the complete pUB110-protease construct was recovered with 100 % coverage and 100 % nucleotide identity. The complete pUB110-protease was recovered in six out of ten Illumina datasets (60 %). In the remaining four datasets, the construct was recovered with 99.99 % coverage and 99.99 % identity. Furthermore, all GMM junctions associated with the pUB110-protease GMM were detected with 100 % identity and coverage in all datasets (Illumina, ONT and hybrid).

In addition, AMRFinder+ identified *bleO* and *aadD1* in all GMM samples with 100 % identity and coverage. Both genes are part of the GMM construct and their presence further supports the presence of the GMM construct in the isolates. In contrast, the *satA* gene was detected as an imperfect hit (i.e., perfect coverage, but with <100 % identity) (Table 3). This gene is associated with resistance to streptothricin and frequently occurs in *wt Bacillus* strains, as shown in the CARD ontology (Alcock et al., 2023). Therefore, the presence of this gene is not informative and does not indicate transgenic modification. Hence, this gene is not included in the GMM marker database (Table 1). Overall, the performance of the AMR and GMM detection assays was perfect, which correctly identified the construct and the associated AMR markers (Table 1) in the GMM strains (Table 3) as well as the AMR genes naturally present in the *wt* strains.

3.1.1.3. GMM strain detection. For all GMM samples and all sequencing technologies, the workflow identified the correct strain cluster (representative: "Baci_velezensis_10075") associated with the known GMM. As both the GMM construct and the correct strain cluster were identified in all GMM samples, the workflow correctly reported a match to the pUB110-protease GMM from the database for all ten datasets and for all sequencing technologies.

3.1.1.4. Other assays. The workflow did not detect any genes from the VFDB in any of the datasets. The Rep1 replicon from the PlasmidFinder database was detected in all assemblies, as expected. In addition, for the ONT and hybrid samples, MOB-Suite identified a complete 6.7 kb plasmid, closely matching the length of the pUB110-protease construct in five samples (samples cob92, crystal1, pilsner12, pilsner22 and pure1). In the remaining five ONT-only and hybrid assemblies, only parts of the plasmid could be reconstructed by MOB-suite. Nevertheless, MOB-suite inferred for the plasmidic contigs the same plasmid primary cluster identifier as the pUB110-protease construct (i.e., AC670). Interestingly, the genomic context predicted that both AMR-conferring genes *bleO* and *aadD1* were located on this MGE for all GMM samples, further supporting the identification of this MGE as the expected pUB110-protease construct. In contrast, the Illumina short-read only assemblies did not contain (near) full-length plasmid contigs, and MOB-suite was unable to assign a plasmid primary cluster to these contigs. In addition, the rMLST assay identified all samples as *Bacillus* sp. for all sequencing technologies, and the MLST typing assays classified all samples as ST140. Finally, the closest genomes identified by the FastANI assay for all assemblies were *B. velezensis* strains, consistent with the species assignment of the StrainGST assay (see reports hosted on Zenodo: https://doi.org/10.5281/zenodo.12819283).

3.1.2. wt B. velezensis strains

3.1.2.1. QC checks. All *wt* samples passed the QC checks, except for sample LMG22478 (*B. velezensis*) which failed the percentage of MLST loci detected check (i.e., minimum 90 % of alleles detected). This was due to a novel allele sequence that was not present yet in the PubMLST database. The novel allele was submitted to the PubMLST database and the dataset was retained for further analysis.

3.1.2.2. GMM construct and GMM strain detection. No GMM-associated elements were detected in any of the six *wt* strains, i.e., no markers, constructs or junctions from the GMM databases were detected, nor the strain associated with a known GMM ("Baci_velezensis_10075"). Regarding the presence of AMR-associated genes, all wild-type strains except LMG26770 carried the *satA* gene, and three strains also carried the *clbA* gene, which is associated with resistance to lincosamides, macrolides, and streptogramins. These two genes are commonly observed in the *Bacillus* genus, as shown in the CARD ontology pages for those genes (Alcock et al., 2023). Results for the other assays are available in the HTML reports hosted on Zenodo (https://doi.org/10.5281/zenodo.12819283).

3.1.3. Application of the workflow to the pGM-rib GMM

As an independent test case, we used the workflow to analyze a single replicate of a strain carrying the pGM-rib GMM construct (i.e., a 38.6 kb plasmid modified to overproduce riboflavin). Both Illumina and ONT data were available for this strain. The full results can be found in the Supplementary Material (Section S1). In summary, the data quality was deemed sufficient for subsequent analysis, as all samples passed the quality checks. Table S5 shows the results of the alignment of the assembled contigs against the Blast+ nt database. This alignment showed that some reads, which were labeled as *Escherichia* contamination, were mapped to the *Escherichia* phage Lambda sequence, indicating that the failed QC check was unlikely due to true contamination. Table S6 shows the results of aligning the assembled contigs against the pGM-rib sequence, indicating that the complete construct is present, albeit with an inversion at the start of the construct, likely due to an assembly error. Additionally, the correct strain were identified in the Illumina-only, ONT-only, and hybrid datasets, indicating that the workflow could correctly characterize these datasets. Furthermore, additional assays such as the detection of AMR genes, detection of transgenic junctions, and the detected of plasmid replicons yielded the expected results.

3.2. Performance evaluation – Bacillus cereus

Secondly, we evaluated the workflow for the characterization of *B. cereus*. As *B. cereus* contamination is often associated with cases of food poisoning, it is particularly important to detect the presence of

genes associated with virulence or AMR. To evaluate this, we applied the workflow to 66 *B. cereus s.l.* samples, consisting of nine in-house isolates collected from enzymatic solutions (Illumina only) and 57 publicly available *B. cereus s.s.* biovar *Thurigiensis* isolates, for which both Illumina and ONT sequencing data were available. We validated the species assignment for all 66 samples using an independent clustering-based species identification approach (Fig. S2-S3, Table S11). Note that AMR and virulence gene detection was performed on these strains, but performance evaluation was not possible due to the lack of reference data for these assays.

3.2.1. QC checks

In total, 65 (98.5 %) Illumina-only (Table S7), 51 (89.4 %) ONT-only (Table S8) and 41 (71.9 %) hybrid samples (Table S9) passed the QC checks. The Illumina sample that failed the QC checks was flagged as contaminated by ConFindr (sample PS00434). Of the six ONT-only datasets that failed the QC checks, two were flagged by Kraken2 as contaminated. However, in both cases, the contaminant species were Bacilli phages (*Camtrevirus* and *Betatectivirus*) and we therefore retained these samples. For three of the remaining ONT samples that failed QC checks, the percentage of complete BUSCO genes in the assembly (72.6 %, 69.4 % and 72.6 %) was below the threshold. However, the addition of short reads in the hybrid assembly improved the overall assembly quality, increasing the number of BUSCO genes identified above the threshold (Table S1). Therefore, we decided to also retain these three samples (both Illumina and ONT sequencing). Finally, the remaining sample that failed the QC checks did not have sufficient coverage to generate a complete assembly and was therefore not retained (sample PS00122).

As both ONT and Illumina sequencing data were available for the 57 publicly available samples, the hybrid approach could also be evaluated (Wick et al., 2023). Of the hybrid datasets, 16 failed the QC checks, the majority of which (12/16, 75 %) did not pass due to the mapping rate of the Illumina reads to the assembled contigs, which was below the failure threshold (Table S1). Nevertheless, these datasets were retained, as all the other QC metrics indicated that the assembly was of high quality. Two out of the remaining four samples that failed the QC checks were retained (both Illumina and ONT), as they consisted of the two Bacilli phage contaminations identified by Kraken2 (identical contamination as those mentioned above for ONT-only samples). The third sample did not have sufficient coverage (PS00122, identical sample to the ONT one mentioned above) and the fourth sample was flagged as contaminated by ConFindr (PS00434, identical sample to the Illumina one mentioned above).

In summary, 64 of the 66 samples were retained for the performance evaluation with Illumina and ONT sequencing (when available). The removed samples are sample PS00122 (low coverage in the ONT data) and sample PS00434 (suspected inter-species contamination in the Illumina data). Note that all data for these two samples were discarded, even if the quality was sufficient for one of the sequencing technologies.

3.2.2. Species identification

Sub-species identification is automated in the BTyper3 tool included in our bioinformatics workflow, and is performed by ANI with an updated nomenclature (Carroll, Wiedmann, & Kovac, 2020). Notably, in the BTyper3 database, *B. mosaicus* consists of a larger group, which includes the previously named *B. paranthracis* species, among others. This updated nomenclature has not yet been universally adopted, and RefSeq still uses the conventional taxonomic labels (O'Leary et al., 2016). To validate the subspecies identification of our bioinformatics workflow, a set of nine in-house *B. cereus s.l.* samples were collected from enzymatic solutions and species labels were determined by an independent bioinformatics approach (see section 2.2.1.2 and Fig. S2-S3). The subspecies prediction of our bioinformatics workflow, using BTyper3, yielded perfectly matching results with these nine Illumina sequencing data (Table 4). Of note, some samples had unresolved species assignment in the clustering approach between *B. luti* and *B. mosaicus*. This result can be attributed to the phylogenetic relationships between those two genomospecies since it was recently observed that the *B. luti* clade is branching as a separate lineage within the *B. mosaicus* clade (Carroll, Cheng, et al. 2020). Furthermore, we also predicted species assignment for all *B. cereus s.s.* biovar *Thurigiensis*, which resulted in a perfectly concordant species assignment by the clustering approach and by our workflow for all samples included (Table S10).

Table 4
Comparison between the predicted and expected species for *B. cereus s.l.* data.

Isolate	Predicted species (BTyper3)	Expected species (clustering approach)
S28	*B. mosaicus*	*B. mosaicus, B. luti*
S29	*B. cereus s.s.*	*B. cereus s.s.*
S30	*B. cereus s.s.*	*B. cereus s.s.*
S37	*B. mosaicus*	*B. mosaicus, B. luti*
S31	*B. cereus s.s.*	*B. cereus s.s.*
S32	*B. mosaicus*	*B. mosaicus, B. luti*
S33	*B. mosaicus*	*B. mosaicus, B. luti*
S34	*B. toyonensis*	*B. toyonensis*
S35	*B. toyonensis*	*B. toyonensis*

This table shows for each isolate (first column) the species detected by our workflow (using BTyper3) in the second column and the expected species as determined by the independent clustering approach in the third column. Accession numbers are available in Table S12.

An additional species typing performed by BTyper3 is the *panC* sequence typing. For the *B. cereus s.s.* biovar *Thurigiensis* samples, the predicted *panC* group assignments by BTyper3 were reported in the original study, and could hence be compared to our workflow. Here, the workflow identified few discordant BTyper3 *panC* species assignments compared to the original study. Four Illumina-only samples (7.2 %), nine ONT-only samples (16.4 %), and four hybrid sequencing datasets (7.2 %) had a different phylogenetic group assigned compared to the original study (Table S10). However, the clustering-based species assignment validated the BTyper3 species assignment, supporting the results provided by our workflow.

3.2.3. AMR and virulence genes detection

The AMRFinder+ assay detected four different perfect matches to AMR-associated genes (Table S10). The *fosB* gene and the *fosBx1* gene, both associated with fosfomycin resistance, were detected in three and five samples respectively. Additionally, the *qacH* gene, which encodes a subunit of the QAC multidrug efflux pump, was detected in five samples. Finally, The *bla2* gene was detected in two samples. Three of those genes are frequently found in *Bacillus cereus s.l.* (*fosB, fosBx1, bla2*) and *qacH* is frequently found in *Staphylococcus saprophyticus*, as shown in the CARD ontology pages for those genes (Alcock et al., 2023). Further investigation showed that this gene was detected on plasmidic contigs, which were not inferred as plasmidic by the MOB-suite assay, as this plasmid sequence is not present in the MOB-suite database. These results were generally consistent across sequencing technologies when both datasets were available, where the genes were detected in at least two different assemblies (4/9 in the three assemblies, 3/9 in Illumina and hybrid assemblies, and 2/9 in ONT and hybrid assemblies).

Among imperfect hits, 14 different AMR-associated genes were identified. Multiple genomes had imperfect matches to AMR-associated genes frequently found in Bacilli, such as *fosB, qacH, bla2,* but also *satA*, *mphM* and *mphL* (associated with macrolide resistance). The AMRFinder+ assay also revealed the presence of genes infrequently or rarely found in Bacilli genomes, such as *tet(45)* (associated with resistance to tetracycline), *vanS-Pt* (associated with resistance to vancomycin) and *vanR-A* (associated with resistance to glycopeptide). Interestingly, a recent study reported the presence of these three genes in *B. cereus s.s.* strains isolated from Galantamine, a dietary supplement available in the US (Cohen et al., 2024).

Various virulence genes were also identified by the workflow using

BTyper3 (Table S10). Of note, the three genes encoding the three sub-units of the non-haemolytic enterotoxin (*nheA*, *nheB*, *nheC*), involved in diarrhea were identified in all but two samples (ONT-only assemblies of PS00536 and PS00547) where only two sub-units were identified. Furthermore, the VirulenceFinder assay provides additional metrics to the BTyper3 assay, such as identity and coverage to reference genes, that are not reported by BTyper3. For example, the VirulenceFinder assay found that in seven *B. cereus s.l.* samples the coverage of all three genes was 100 % and the identity was over 90 %. In the remaining two *B. cereus s.l.* samples (S34 and S35), the *nheC* gene was covered for 99.1 % with a nucleotide an identity of 90.65 %.

3.2.4. *Bt* genes detection

In the original paper, the authors detected *Bt* genes in 18/55 *B. cereus s.s.* Illumina-only assemblies (32.7 %) and in 17/55 hybrid assemblies (30.9 %) using BTyper3 (Chung et al., 2024). Overall, our results are consistent with the previously reported results (Table S10). The workflow correctly detected the presence of *Bt* genes in 18/18 Illumina-only assemblies (100 %) and in 14/17 hybrid assemblies (82.3 %).

For the remaining 37 *B. cereus s.s.* Illumina assemblies where the authors did not detect *Bt* genes with BTyper3, the current workflow detected *Bt* genes in a single assembly, for which phenotypic production of *Bt* genes had been observed (PS00446, 2.7 %), and none in any of the other datasets. Finally, for the remaining 38 hybrid assemblies, five samples were detected as carrying *Bt* genes (5/38, 13.1 %).

4. Discussion & conclusion

In the context of the food and feed chain, WGS has gained prominence as a sensitive, powerful and accurate method for detecting and characterizing bacterial pathogen isolates (Carroll et al., 2019). In this study, we have applied WGS for the characterization of viable GM *B. subtilis s.l.* strains. Identification and characterization of viable GM strains in the food chain is crucial for enforcement laboratories, as the use of GMM is heavily regulated and the presence of viable bacteria in the final product is strictly unauthorized on the EU market. Secondly, the workflow can perform the identification and characterization of potentially pathogenic *B. cereus s.l.* strains in food and feed products, given that *B. cereus* contamination poses a significant public health risk. To address these challenges, our bioinformatics methodology is implemented as an automated, end-to-end, bioinformatics workflow that is compatible with Illumina, ONT and hybrid sequencing data, and is specifically designed for easy adoption by enforcement laboratories.

For *Bacillus subtilis*, the GMM detection is performed using a two-step approach: (1) screening for known GMM constructs and (2) strain identification at single nucleotide polymorphism (SNP) resolution. In contrast to plants, where detecting the construct alone is sufficient to identify a GMO, GMM detection requires not only construct identification but also precise strain-level characterization (D'aes et al., 2021). The workflow will only report an exact match if both the construct and the strain match a known GMM in the database. Notably, the bioinformatic approach differs from the wet lab approach, which relies on a sequential search for signatures of GMM elements and transgenic junctions. First, qPCR assays detect suspected GM elements, such as AMR-associated genes and plasmid vectors. If positive matches are identified, the junctions are then targeted to circumvent the exact GM construct that was used. Although our approach relies on the detection of complete constructs, we have included the detection of these junctions as part of the workflow (Table 2).

The performance of this approach was validated using ten positive (i.e., GM *B. velezensis*) and six negative control (i.e., *wt B. velezensis*) datasets, all of which were correctly classified by the workflow as either GM or *wt*, regardless of the sequencing technology used (i.e., Illumina, ONT or hybrid). Interestingly, the strain assignments obtained using ONT data matched those obtained using Illumina data perfectly, even though StrainGST does not explicitly support ONT input. This suggests that the k-mer-based approach used by StrainGST is not significantly affected by the generally lower quality of ONT reads compared to Illumina reads. However, further testing is needed to confirm whether this remains true for other strains and datasets. Moreover, we demonstrated the applicability of the workflow to another GM construct (i.e., the pGM-rib construct), which was matched to the correct combination of GM construct and host strain.

The main limitation of the GMM detection is the fact that it relies on a database of known GMM strains and their corresponding constructs. Currently, this database only contains two entries, since only two genomically characterized GMM strains have been identified and documented in the literature so far (Berbers et al., 2020; D'aes et al., 2021; Fraiture, Bogaerts, et al., 2020). This hinders real-world application and generalizability, as it cannot identify unknown or novel GM strains. However, our aim is to expand this database progressively as more GMMs are isolated and characterized, which will require substantial effort from enforcement laboratories since the structure of GM constructs is often unknown (Fraiture, Bogaerts, et al., 2020). To effectively support surveillance efforts such as the workflow described in this study, the availability of a comprehensive public database containing known GMM events for the food and feed chain and their complete genomic information would be highly beneficial. Nevertheless, for GM strains not included in the database, the assays in the workflow still provide valuable information, enabling end users to make informed assessments of the GM status of a strain. In particular, the provided GMM databases contain genomic elements that are often associated with GM Bacilli, including AMR-associated genes used as selection markers and plasmids used as a backbone for genetic modifications. These elements are always reported if present, even if the combination of GM markers and identified strain does not correspond to a GMM in the database. Furthermore, variants of these markers are also be reported if they meet the filtering criteria. This enables the end user to determine the necessary experimental validation strategy to confirm a potential novel GMM. Currently, this still requires specialized domain knowledge, for example, to differentiate between AMR genes that are intrinsic or naturally acquired and those introduced through genetic modification (Fraiture, Deckers, et al., 2020).

Accurate identification of plasmids is essential for detecting plasmid-encoded GMMs and assessing the potential risks posed by pathogenic *B. cereus*. This makes plasmid reconstruction and characterization critical steps in the workflow. Although MOB-suite can theoretically group contigs from the same plasmid, this process can be error-prone, particularly for fragmented assemblies from short-read sequencing data (Beh et al., 2025; Robertson et al., 2020). Due to the longer read lengths, assembly fragmentation is expected to be lower for long-read or hybrid datasets. However, we were unable to obtain complete circular chromosome sequences for any of the validation datasets (Table S4), due to the relatively short ONT reads. This was limitation was also noted in the original publication (D'aes et al., 2021), and various benchmarking studies have shown that complete, unfragmented assemblies are often not obtained, even with hybrid sequencing data (Chen et al., 2020; De Maio et al., 2019). To minimize the effect of the assembly step on the detection of GMM constructs, the workflow uses a read-mapping-based approach (i.e., KMA), which is more sensitive, especially for low coverage targets, such as those on low copy number plasmids. In the performance evaluation, we have demonstrated that the this approach enabled the complete detection of the constructs. The thresholds for identifying constructs were adapted from the EFSA guidelines for reporting elements of concern (European Food Safety Authority (EFSA), 2024).

The workflow has been developed largely in accordance with the EFSA requirements for WGS analysis of micro-organisms intentionally used in the food chain (EFSA 2024), with some adaptations to enable automated accurate detection of known GMMs. While EFSA recommends genome-wide comparisons with *wt* references to identify any genomic variation, whether natural or unnatural, this approach is

difficult to automate and may lead to ambiguous results due to natural genomic variability, such as HGT. Therefore, we have opted for a targeted approach that differs from the strategy proposed in the EFSA guidelines. This approach provides clear and interpretable results (i.e., a strain is or is not a GMM) and is optimized for practical use by enforcement laboratories. Secondly, the workflow uses only a single database for the detection of AMR-related genes, instead of using (at least) two different ones as stated in the guidelines (European Food Safety Authority (EFSA), 2024). This strategy was followed to reduce redundancy in the output and to facilitate interpretation. Lastly, we included several additional QC checks, adapted from extensively validated workflows for other bacterial species (Bogaerts et al., 2019; Bogaerts, Delcourt, et al., 2021; Bogaerts, Nouws, et al., 2021).

For *B. cereus s.l.*, the workflow focuses on subspecies identification and on assessing the pathogenic potential of strains by characterizing the genes encoding virulence factors and AMR-associated genes. The workflow also includes typing methods, such as rMLST and cgMLST, which can be used for constructing phylogenies to study the relationships between isolates (Tourasse et al., 2023). This can be crucial, for example, in tracing the origin of foodborne outbreaks (Bogaerts et al., 2023). The performance of the workflow was assessed by re-analyzing a previously published dataset of *B. cereus s.s.* biovar *Thurigiensis* and evaluating the concordance of (1) species identification and (2) identification of *Bt* genes. The species identification part was also validated by independently sequencing nine additional *B. cereus s.l.* and applying an independent clustering approach to assign subspecies labels (see Section 2.2.1.2 and Figs. S2–S3). While our results were mostly in line with the published results, we observed some differences regarding the detection of *Bt* genes and species identification. The previous study reported only binary information on the presence or absence of *Bt* genes, without detailed gene-level information, limiting direct comparison (Chung et al., 2024). Discrepancies in genomic detection between the original study and our workflow may be related to database content, as *Bt* genes include numerous and diverse gene categories, some of which may not have been fully characterized yet (Panneerselvam et al., 2022). Nevertheless, our workflow improved the *Bt* genes detection for at least one sample where the hybrid assembly approach allowed the detection of a *Bt* gene that was not detected in the ONT-only assembly. This result highlights the added value of using hybrid assembly procedures to improve assembly quality and gene detection accuracy (Wick et al., 2023). In addition, species assignments were largely consistent with the original study. The observed differences in our workflow compared to the original study were supported by the independent clustering approach. These results suggest that our workflow is suited for the identification and characterization of *B. cereus. s.l.* isolates. A limitation of our performance evaluation is that the presence of AMR-associated genes and of the other toxin-producing genes could not be validated, as reference information for these assays was not available in the validation dataset. Moreover, the presence of a gene does not necessarily translate into the corresponding phenotype. However, the approach itself has been extensively validated in several species, generally showing high concordance between predicted and observed phenotypes (; Bogaerts et al., 2019; Feldgarden et al., 2019). In the context of risk assessment, this workflow can therefore serve as an extensive initial screening tool, although it may require additional follow-up experiments. For example, predicted AMR or toxigenicity phenotypes may require confirmation through in vivo testing, or verification of transgenic construct or junctions with qPCR. Alternatively, laboratories could conduct comprehensive validation of the workflow using datasets that include molecular testing results and are representative of the intended application.

Our workflow complements and extends other existing software for the detection of (potentially) pathogenic *B. cereus* strains and GM *B. subtilis*. To the best of our knowledge, there are two workflows for the detection of GM bacteria based on WGS data: DUGMO (Hurel et al., 2020) and Synsor (Tay et al., 2024). DUGMO starts by assembling Illumina reads, followed by inferring coding sequences (CDS) from the resulting assembly, which are then compared to the pan-genome of the target species. A machine learning approach is then used to discriminate between host genome CDSs and GM CDSs, for which a database is preemptively built. However, unlike our workflow, DUGMO does not provide a full characterization of the GM strain (i.e., neither the GM construct nor the GM strain), and requires a high-quality pan-genome and a specifically trained model to accurately infer exogenous DNA material. However, in theory, DUGMO is generally applicable, regardless of the target species, whereas our workflow focuses on GM Bacilli. Interestingly, the authors of the DUGMO paper applied the tool to the pGM-rib construct and retrieved the principal genes of interest (i.e., the riboflavin, *cat* and *recA* genes). Synsor is another recently released tool for identifying engineered DNA sequences using an alignment-free approach (Tay et al., 2024). We applied the tool to the GMM datasets used in this study, but it did not classify any of them as GM (results not shown). Therefore, Synsor may not be a suitable alternative to our workflow. For plasmids, alternative strategies based on k-mer signatures have been proposed to detect artificial vector sequences used in genetic engineering (Allen et al., 2008).

Although the presented workflow is tailored towards Bacilli, many of the steps are species-agnostic and have been adapted from workflows previously described for other organisms (Bogaerts et al., 2019; Bogaerts, Nouws, et al., 2021). Pre-processing steps and assays such as read trimming, de novo assembly, and sequence typing have been described previously, and are being used in extensively validated workflows. Similar workflows have been developed and validated by other laboratories, including (Hung et al., 2025; Kohl et al., 2018; Ortega-Sanz et al., 2023; Petit & Read, 2018; Smedile et al., 2025). However, these workflows have only been validated for Illumina input data. To our knowledge, no end-to-end workflows have been validated for bacterial isolate characterization on ONT data, which hinders its integration into routine practice.

For *B. cereus*, BTyper3 is a widely used workflow for the characterization of *B. cereus s.l.* (Carroll, Cheng, et al. 2020). Some alternative tools have been developed specifically for the characterization of *B. cereus s.s.* biovar *Thurigiensis* strains, such as IDOPS (Díaz-Valerio et al., 2021), BtToxin_Digger (Liu et al., 2021) and cry_processor (Shikov et al., 2020), all three focusing on the accurate detection of *Bt* genes. Our workflow uses BTyper3 to characterize *B. cereus* sequencing data, as this tool is widely used and has extensive functionality (Carroll, Cheng, & Kovac, 2020). In addition to the BTyper3 characterization, our workflow integrates several additional assays such as pre-processing (including multiple assembly approaches), validated quality checks, cgMLST profiling and detection of AMR genes and mobile genetic elements.

While our workflow is currently designed for isolate analysis, it could theoretically also be applied to meta-genomes obtained from metagenomics datasets (i.e., by sequencing the input matrix directly). This may increase the number of use cases for the workflow, as metagenomic sequencing does not require isolation or cultivation (Handelsman et al., 1998; Kellenberger, 2001; Ko et al., 2022; Quince et al., 2017). Still, additional assays would be required to validate this approach, given that the current version of the workflow was developed for high quality data from bacterial isolates. A potential strategy would be to first assemble the metagenomic data and then perform binning in order to assign species labels to each bin. The metagenome-assembled genomes could then be fed into the current workflow.

In conclusion, we have developed and validated a novel WGS-based workflow for the comprehensive characterization of *B. cereus s.l.* and *B. subtilis s.l.* isolates. We have demonstrated its performance for accurate detection of GM *B. subtilis* strains, as well as the identification of toxin-encoding genes and correct subspecies identification for *B. cereus* isolates. This workflow can be a valuable complement or alternative to existing molecular biology-based methods, such as qPCR. This can be particularly relevant for enforcement laboratories, competent authorities and industry, as it is a reliable and reproducible method for

detecting unauthorized or potentially hazardous bacterial contaminations (Barbau-Piednoir et al., 2015; Paracchini et al., 2017). The results are presented in an easy-to-interpret format, enabling users, including non-bioinformatics experts, to easily evaluate the risk associated with potential contaminants in food and feed products. To the best of our knowledge, this study provides the first proof of concept for the identification and complete reconstruction of known GMMs, supporting short and/or long read WGS data.

CRediT authorship contribution statement

Maxime Godfroid: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis. **Alexander Van Uffelen:** Writing – review & editing, Software, Methodology, Investigation, Formal analysis. **Marie-Alice Fraiture:** Writing – review & editing, Resources, Methodology, Investigation, Conceptualization. **Sigrid C.J. De Keersmaecker:** Writing – review & editing, Resources, Methodology. **Kevin Vanneste:** Writing – review & editing, Resources, Methodology, Conceptualization. **Nancy H.C. Roosens:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Bert Bogaerts:** Writing – review & editing, Writing – original draft, Supervision, Software, Project administration, Methodology, Formal analysis, Conceptualization.

Funding

This work was supported by the Transversal activities in Applied Genomics Service from Sciensano (Belgium).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank the technicians of the service Transversal activities in Applied Genomics at Sciensano, Belgium for performing the Illumina whole-genome sequencing runs.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fochms.2025.100338.

Data availability

The WGS datasets for all samples used in the performance evaluation are publicly available on SRA (Table S12). The workflow is available for non-commercial use on the Galaxy instance of our institute at https://galaxy.sciensano.be (registration required). HTML reports for the GMM and *B. cereus* analyses, as well as the GMM markers and junctions databases are available on Zenodo: doi.org/10.5281/zenodo.12819283.

References

Alcock, B. P., Huynh, W., Chalil, R., Smith, K. W., Raphenya, A. R., Wlodarski, M. A., , … Tsang, K. K., et al. (2023). CARD 2023: Expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Research, 51*, D690–D699.

Allen, J. E., Gardner, S. N., & Slezak, T. R. (2008). DNA signatures for detecting genetic engineering in bacteria. *Genome Biology, 9*, R56.

Arnold, B. J., Huang, I.-T., & Hanage, W. P. (2022). Horizontal gene transfer and adaptive evolution in bacteria. *Nature Reviews. Microbiology, 20*, 206–218.

Barbau-Piednoir, E., De Keersmaecker, S. C. J., Wuyts, V., Gau, C., Pirovano, W., Costessi, A., … Roosens, N. H. (2015). Genome sequence of EU-unauthorized genetically modified *Bacillus subtilis* strain 2014-3557 overproducing riboflavin, isolated from a vitamin B2 80% feed additive. *Genome Announcements, 3*. https://doi.org/10.1128/genomea.00214-15

Beh, J. Q., Wick, R. R., Howden, B. P., Connor, C. H., & Webb, J. R. (2025). Challenges and considerations for whole-genome-based antimicrobial resistance plasmid investigations. *Antimicrobial Agents and Chemotherapy, 0*. e01097–25.

Bennett, S. D., Walsh, K. A., & Gould, L. H. (2013). Foodborne disease outbreaks caused by *Bacillus cereus*, *Clostridium perfringens*, and *Staphylococcus aureus*—United States, 1998–2008. *Clinical Infectious Diseases, 57*, 425–433.

Berbers, B., Saltykova, A., Garcia-Graells, C., Philipp, P., Arella, F., Marchal, K., … De Keersmaecker, S. C. J. (2020). Combining short and long read sequencing to characterize antimicrobial resistance genes on plasmids applied to an unauthorized genetically modified Bacillus. *Scientific Reports, 10*, 4310.

Bogaerts, B., Delcourt, T., Soetaert, K., Boarbi, S., Ceyssens, P.-J., Winand, R., , … Marchal, K., et al. (2021). A bioinformatics whole-genome sequencing workflow for clinical *Mycobacterium tuberculosis* complex isolate analysis, validated using a reference collection extensively characterized with conventional methods and in silico approaches. *Journal of Clinical Microbiology, 59*. https://doi.org/10.1128/jcm.00202-21

Bogaerts, B., Fraiture, M.-A., Huwaert, A., Van Nieuwenhuysen, T., Jacobs, B., Van Hoorde, K., … Vanneste, K. (2023). Retrospective surveillance of viable *Bacillus cereus* group contaminations in commercial food and feed vitamin B2 products sold on the Belgian market using whole-genome sequencing. *Frontiers in Microbiology, 14*, Article 1173594.

Bogaerts, B., Nouws, S., Verhaegen, B., Denayer, S., Van Braekel, J., Winand, R., , … Marchal, K., et al. (2021). Validation strategy of a bioinformatics whole genome sequencing workflow for Shiga toxin-producing *Escherichia coli* using a reference collection extensively characterized with conventional methods. *Microbial Genomics, 7*, Article 000531.

Bogaerts, B., Van Braekel, J., Van Uffelen, A., D'aes, J., Godfroid, M., Delcourt, T., , … De Keersmaecker, S. C. J., et al. (2025). Galaxy @Sciensano: A comprehensive bioinformatics portal for genomics-based microbial typing, characterization, and outbreak detection. *BMC Genomics, 26*, 20.

Bogaerts, B., Winand, R., Fu, Q., Van Braekel, J., Ceyssens, P.-J., Mattheus, W., … Vanneste, K. (2019). Validation of a bioinformatics workflow for routine analysis of whole-genome sequencing data and related challenges for pathogen typing in a European National Reference Center: *Neisseria meningitidis* as a proof-of-concept. *Frontiers in Microbiology, 10*, 362.

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

Carattoli, A., Zankari, E., García-Fernández, A., Voldby Larsen, M., Lund, O., Villa, L., … Hasman, H. (2014). In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrobial Agents and Chemotherapy, 58*, 3895–3903.

Carroll, L. M., Cheng, R. A., & Kovac, J. (2020). No assembly required: Using BTyper3 to assess the congruency of a proposed taxonomic framework for the *Bacillus cereus* group with historical typing methods. *Frontiers in Microbiology, 11*, Article 580691.

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

Carroll, L. M., Wiedmann, M., Mukherjee, M., Nicholas, D. C., Mingle, L. A., Dumas, N. B., … Kovac, J. (2019). Characterization of emetic and diarrheal *Bacillus cereus* strains from a 2016 foodborne outbreak using whole-genome sequencing: Addressing the microbiological, epidemiological, and bioinformatic challenges. *Frontiers in Microbiology, 10*, 144.

Chen, S. (2023). Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta, 2*, Article e107.

Chen, Z., Erickson, D. L., & Meng, J. (2020). Benchmarking hybrid assembly approaches for genomic analyses of bacterial pathogens using Illumina and Oxford Nanopore sequencing. *BMC Genomics, 21*, 631.

Chung, T., Salazar, A., Harm, G., Johler, S., Carroll, L. M., & Kovac, J. (2024). Comparison of the performance of multiple whole-genome sequence-based tools for the identification of *Bacillus cereus* sensu stricto biovar thuringiensis. *Applied and Environmental Microbiology, 90*. e01778–23.

Clausen, P. T. L. C., Aarestrup, F. M., & Lund, O. (2018). Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics, 19*, 307.

Cohen, P. A., Jacobs, B., Van Hoorde, K., & Vanhee, C. (2024). Accuracy of labeling of galantamine generic drugs and dietary supplements. *JAMA, 331*, 974–976.

D'aes, J., Fraiture, M.-A., Bogaerts, B., De Keersmaecker, S. C. J., Roosens, N. H. C., & Vanneste, K. (2021). Characterization of genetically modified microorganisms using short- and long-read whole-genome sequencing reveals contaminations of related origin in multiple commercial food enzyme products. *Foods, 10*, 2637.

D'aes, J., Fraiture, M.-A., Bogaerts, B., De Keersmaecker, S. C. J., Roosens, N. H. C. J., & Vanneste, K. (2022). Metagenomic characterization of multiple genetically modified *Bacillus* contaminations in commercial microbial fermentation products. *Life, 12*, 1971.

De Coster, W., & Rademakers, R. (2023). NanoPack2: Population-scale evaluation of long-read sequencing data. *Bioinformatics, 39*, btad311.

De Maio, N., Shaw, L. P., Hubbard, A., George, S., Sanderson, N. D., Swann, J., , … Hoosdally, S. J., et al. (2019). Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microbial Genomics, 5*, Article e000294.

Deckers, M., Deforce, D., Fraiture, M.-A., & Roosens, N. H. C. (2020). Genetically modified micro-organisms for industrial food enzyme production: An overview. *Foods, 9*, 326.
Díaz-Valerio, S., Lev Hacohen, A., Schöppe, R., & Liesegang, H. (2021). IDOPS, a profile HMM-based tool to detect pesticidal sequences and compare their genetic context. *Frontiers in Microbiology, 12*, Article 664476.
van Dijk, L. R., Walker, B. J., Straub, T. J., Worby, C. J., Grote, A., Schreiber, H. L., , … Manson, A. L., et al. (2022). StrainGE: A toolkit to track and characterize low-abundance strains in complex microbial communities. *Genome Biology, 23*, 74.
Ehling-Schulz, M., Lereclus, D., & Koehler, T. M. (2019). The *Bacillus cereus* group: Bacillus species with pathogenic potential. *Microbiology Spectrum, 7*. https://doi.org/10.1128/microbiolspec.gpp3-0032–2018
European Food Safety Authority (EFSA). (2024). EFSA statement on the requirements for whole genome sequence analysis of microorganisms intentionally used in the food chain. *EFSA Journal, 22*, Article e8912.
Feldgarden, M., Brover, V., Gonzalez-Escalona, N., Frye, J. G., Haendiges, J., Haft, D. H., , … Tillman, G. E., et al. (2021). AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Scientific Reports, 11*, 12728.
Feldgarden, M., Brover, V., Haft, D. H., Prasad, A. B., Slotta, D. J., Tolstoy, I., , … McDermott, P. F., et al. (2019). Validating the AMRFinder Tool and Resistance Gene Database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrobial Agents and Chemotherapy, 63*. https://doi.org/10.1128/aac.00483-19
Fraiture, M.-A., Bogaerts, B., Winand, R., Deckers, M., Papazova, N., Vanneste, K., … Roosens, N. H. C. (2020). Identification of an unauthorized genetically modified bacteria in food enzyme through whole-genome sequencing. *Scientific Reports, 10*, 7094.
Fraiture, M.-A., Deckers, M., Papazova, N., & Roosens, N. H. C. (2020). Are antimicrobial resistance genes key targets to detect genetically modified microorganisms in fermentation products? *International Journal of Food Microbiology, 331*, Article 108749.
Fraiture, M.-A., Gobbo, A., Guillitte, C., Marchesi, U., Verginelli, D., De Greve, J., … Roosens, N. H. C. (2024). Pilot market surveillance of GMM contaminations in alpha-amylase food enzyme products: A detection strategy strengthened by a newly developed qPCR method targeting a GM *Bacillus licheniformis* producing alpha-amylase. *Food Chemistry: Molecular Sciences, 8*, Article 100186.
Fraiture, M.-A., Gobbo, A., Marchesi, U., Verginelli, D., Papazova, N., & Roosens, N. H. C. (2021). Development of a real-time PCR marker targeting a new unauthorized genetically modified microorganism producing protease identified by DNA walking. *International Journal of Food Microbiology, 354*, Article 109330.
Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chemistry & Biology, 5*, R245–R249.
Hung, H. C. H., Kumar, N., Dyster, V., Yeats, C., Metcalf, B., Li, Y., … Lo, S. W. (2025). GPS pipeline: Portable, scalable genomic pipeline for *Streptococcus pneumoniae* surveillance from global pneumococcal sequencing project. *Nature Communications, 16*, 8345.
Hurel, J., Schbath, S., Bougeard, S., Rolland, M., Petrillo, M., & Touzain, F. (2020). DUGMO: Tool for the detection of unknown genetically modified organisms with high-throughput sequencing data for pure bacterial samples. *BMC Bioinformatics, 21*, 284.
Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications, 9*, 5114.
Jolley, K., Bray, J., & Maiden, M. (2018). *Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. 3*. Available from: https://wellcomeopenresearch.org/articles/3-124/v1.
Jolley, K. A., Bliss, C. M., Bennett, J. S., Bratcher, H. B., Brehony, C., Colles, F. M., , … Cody, A. J., et al. (2012). Ribosomal multilocus sequence typing: Universal characterization of bacteria from domain to strain. *Microbiology, 158*, 1005–1015.
Kellenberger, E. (2001). Exploring the unknown. *EMBO Reports, 2*, 5–7.
Ko, K. K. K., Chng, K. R., & Nagarajan, N. (2022). Metagenomics-enabled microbial surveillance. *Nature Microbiology, 7*, 486–496.
Kohl, T. A., Utpatel, C., Schleusener, V., Filippo, M. R. D., Beckert, P., Cirillo, D. M., & Niemann, S. (2018). MTBseq: A comprehensive pipeline for whole genome sequence analysis of *Mycobacterium tuberculosis* complex isolates. *PeerJ, 6*, Article e5895.
Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology, 37*, 540–546.
Kozyreva, V. K., Truong, C.-L., Greninger, A. L., Crandall, J., Mukhopadhyay, R., & Chaturvedi, V. (2017). Validation and implementation of clinical laboratory improvements act-compliant whole-genome sequencing in the public health microbiology laboratory. *Journal of Clinical Microbiology, 55*, 2502–2520.
Liu, B., Zheng, D., Zhou, S., Chen, L., & Yang, J. (2022). VFDB 2022: A general classification scheme for bacterial virulence factors. *Nucleic Acids Research, 50*, D912–D917.
Liu, H., Zheng, J., Bo, D., Yu, Y., Ye, W., Peng, D., & Sun, M. (2021). BtToxin_Digger: A comprehensive and high-throughput pipeline for mining toxin protein genes from *Bacillus thuringiensis*. *Bioinformatics, 38*, 250–251.
Low, A. J., Koziol, A. G., Manninger, P. A., Blais, B., & Carrillo, C. D. (2019). ConFindr: Rapid detection of intraspecies and cross-species contamination in bacterial whole-genome sequence data. *PeerJ, 7*, Article e6995.
Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). 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, 38*, 4647–4654.
Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., & Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics, 34*, i142–i150.
Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., et al. (2021). Sustainable data analysis with Snakemake. Available from: https://f1000research.com/articles/10-33.
Nanoporetech. (2024). *Medaka*. Oxford Nanopore Technologies. Available from: https://github.com/nanoporetech/medaka.
NCBI. (2023). *ncbi/sra-human-scrubber*. NCBI - National Center for Biotechnology Information/NLM/NIH. Available from: https://github.com/ncbi/sra-human-scrubber.
O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., , … Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research, 44*, D733–D745.
Ortega-Sanz, I., Barbero-Aparicio, J. A., Canepa-Oneto, A., Rovira, J., & Melero, B. (2023). CamPype: An open-source workflow for automated bacterial whole-genome sequencing analysis focused on Campylobacter. *BMC Bioinformatics, 24*, 291.
Panneerselvam, S., Mishra, R., Berry, C., Crickmore, N., & Bonning, B. C. (2022). BPPRC database: A web-based tool to access and analyse bacterial pesticidal proteins. *Database, 2022*, Article baac022.
Paracchini, V., Petrillo, M., Reiting, R., Angers-Loustau, A., Wahler, D., Stolz, A., , … Meinel, D. M., et al. (2017). Molecular characterization of an unauthorized genetically modified *Bacillus subtilis* production strain identified in a vitamin B2 feed additive. *Food Chemistry, 230*, 681–689.
Petit, R. A., & Read, T. D. (2018). *Staphylococcus aureus* viewed from the perspective of 40,000+ genomes. *PeerJ, 6*, Article e5261.
Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020). Using SPAdes De Novo assembler. *Current Protocols in Bioinformatics, 70*, Article e102.
Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology, 35*, 833–844.
Ramnath, V., Larralde, M., Menchik, P., Buehler, A. J., Harrand, A. S., Chung, T., Wei, X., Raghuram, V., Gourlé, H., Pierneef, R., et al. (2023). A community-curated, global atlas of *Bacillus cereus* sensu lato genomes for epidemiological surveillance, 2023.12.20.572685. Available from: https://www.biorxiv.org/content/10.1101/2023.12.20.572685v1.
Robertson, J., Bessonov, K., Schonfeld, J., & Nash, J. H. E. (2020). Universal whole-sequence-based plasmid typing and its utility to prediction of host range and epidemiological surveillance. *Microbial Genomics, 6*, Article e000435.
Robertson, J., & Nash, J. H. E. (2018). MOB-suite: Software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microbial Genomics, 4*, Article e000206.
Sanderson, N. D., Hopkins, K. M. V., Colpus, M., Parker, M., Lipworth, S., Crook, D., & Stoesser, N. (2024). Evaluation of the accuracy of bacterial genome reconstruction with Oxford Nanopore R10.4.1 long-read-only sequencing. *Microbial Genomics, 10*, Article 001246.
Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One, 11*, Article e0163962.
Shen, W., Sipos, B., & Zhao, L. (2024). SeqKit2: A Swiss army knife for sequence and alignment processing. *iMeta, 3*, Article e191.
Sherry, N. L., Horan, K. A., Ballard, S. A., Gonçalves da Silva, A., Gorrie, C. L., Schultz, M. B., , … Stinear, T. P., et al. (2023). An ISO-certified genomics workflow for identification and surveillance of antimicrobial resistance. *Nature Communications, 14*, 60.
Shikov, A. E., Malovichko, Y. V., Skitchenko, R. K., Nizhnikov, A. A., & Antonets, K. S. (2020). No more tears: Mining sequencing data for novel Bt cry toxins with CryProcessor. *Toxins, 12*, 204.
Simon, A. (2010). FastQC A Quality Control tool for High Throughput Sequence Data. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
Smedile, D., Diaconu, E. L., Grelloni, M., Middei, B., Carfora, V., Battisti, A., … Franco, A. (2025). Enteroflow: Automated pipeline for in silico characterization of *Enterococcus faecium*/faecalis isolates from short reads. *International Journal of Molecular Sciences, 26*, 9441.
Tay, A. P., Didi, K., Wickramarachchi, A., Bauer, D. C., Wilson, L. O. W., & Maselko, M. (2024). Synsor: A tool for alignment-free detection of engineered DNA sequences. *Frontiers in Bioengineering and Biotechnology, 12*, Article 1375626.
Tourasse, N. J., Jolley, K. A., Kolstø, A.-B., & Økstad, O. A. (2023). Core genome multilocus sequence typing scheme for *Bacillus cereus* group bacteria. *Research in Microbiology, 174*, Article 104050.
Wang, K., Shu, C., Bravo, A., Soberón, M., Zhang, H., Crickmore, N., & Zhang, J. (2023). Development of an online genome sequence comparison resource for *Bacillus cereus* sensu lato strains using the efficient composition vector method. *Toxins, 15*, 393.
Waskom, M. (2021). seaborn: statistical data visualization. *The Journal of Open Source Software, 6*, 3021.
Wick, R. R., & Holt, K. E. (2022). Polypolish: Short-read polishing of long-read bacterial genome assemblies. *PLoS Computational Biology, 18*, Article e1009802.
Wick, R. R., Judd, L. M., & Holt, K. E. (2023). Assembling the perfect bacterial genome using Oxford Nanopore and Illumina sequencing. *PLoS Computational Biology, 19*, Article e1010905.
Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology, 20*, 257.
Zimin, A. V., & Salzberg, S. L. (2020). The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS Computational Biology, 16*, Article e1007981.