Food Chemistry: Molecular Sciences 11 (2025) 100309

Contents lists available at ScienceDirect

Food Chemistry: Molecular Sciences

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

Detection of *Bacillus* production strains and contaminants in food enzyme products

Alexander Van Uffelen [a,b,c], Andrés Posadas [a,b,c], Marie-Alice Fraiture [a], Nancy H.C. Roosens [a], Sigrid C.J. De Keersmaecker [a], Kathleen Marchal [b,c], Kevin Vanneste [a,*]

[a] *Transversal Activities in Applied Genomics, Sciensano, 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:
Oxford nanopore technologies
Nanopore sequencing
Shotgun metagenomics
Taxonomic classification
Detection threshold
Contamination

ABSTRACT

Shotgun metagenomics enables taxonomic analysis of microbial communities by aligning sequencing reads to reference genomes, for which interpretation of alignment results often lacks standardization and relies on arbitrary abundance thresholds. This can bias species detection, especially for low-abundance or taxonomically complex genera like *Bacillus*, where closely related species may differ in safety and function, and their co-occurrence increases misclassification risk. This study presents a bioinformatics framework for defining detection thresholds of biological contaminations in samples using nanopore shotgun metagenomics data, demonstrated through a case study on *Bacillus subtilis sensu lato* (*s.l.*) and *Bacillus cereus s.l.* contaminations in food enzyme (FE) products. The framework was developed by employing *in silico* mixes of isolate sequencing data of different *B. subtilis* and *B. cereus* species, and uses the tool KMA for taxonomic classification with post-processing steps based on template identity to differentiate true positives from false positives, coupled with curation of the underlying reference genomic database. The performance of the developed framework was afterwards validated with five *in vitro* mixes mimicking potential FE contaminations. Finally, the applicability of the validated framework was evaluated with six real and well-characterized commercial contaminated FE samples, confirming its ability to accurately detect *B. subtilis* and *B. cereus* contaminants, even at low abundances up to a relative abundance of 1 %. In conclusion, we present a bioinformatics framework allowing reliable species-level detection of challenging low-level contaminants in samples using nanopore shotgun metagenomics sequencing, which was successfully applied to identify *B. subtilis* and *B. cereus* contaminations in FE products.

1. Introduction

Shotgun metagenomics allows for the direct sequencing of all extracted DNA in complex samples without the need for isolation and cultivation. This approach enables the identification of species, particularly when used with long-read sequencing platforms (Quince et al., 2017). These platforms can generate fewer but longer genomic fragments, reducing the bioinformatics complexity of taxonomic classification (Kim et al., 2024). Therefore, the combination of shotgun metagenomics with long-read sequencing has proven useful in multiple application domains, such as foodborne outbreak investigation, clinical diagnostics, antibiotic resistance surveillance and phage interactions (Buytaers, Saltykova, et al., 2021; Davis et al., 2023; Taxt et al., 2020; Yahara et al., 2021).

A quintessential step in species detection using shotgun metagenomics is taxonomic classification, whereby raw reads are compared to a reference database to provide a taxonomic annotation per read. The complexity and resource-intensive nature of this problem has led to the development of many classification tools (Breitwieser et al., 2019). Previous studies have performed extensive benchmarking on these different tools to gain insight in their performance on long-read sequencing data (Portik et al., 2022; Simon et al., 2019; Van Uffelen et al., 2024), and identified several challenges in implementing shotgun metagenomics for species detection. Firstly, shotgun metagenomics struggles to distinguish between genuinely present species and those that are incorrectly identified, and is especially prone to the generation

* Corresponding author.
E-mail address: kevin.vanneste@sciensano.be (K. Vanneste).

https://doi.org/10.1016/j.fochms.2025.100309
Received 28 May 2025; Received in revised form 30 September 2025; Accepted 1 October 2025
Available online 4 October 2025

of many false positive species predictions. Such misclassifications can occur due to factors such as high sequence similarity amongst reference genomes (Govender & Eyre, 2022), algorithms prioritizing sensitivity over specificity (Bradford et al., 2024), low-quality reads (Portik et al., 2022), or incomplete and biased databases (Liu et al., 2024). To reduce the chance of propagating false positive species predictions, specific filtering steps such as requiring a minimum number of reads assigned to a species, are often applied before accepting a species to be genuinely present. Secondly, since shotgun metagenomics sequences all genetic material without amplification, it can be challenging to detect DNA of truly present species at low abundances. The limit of detection (LOD) is however still poorly understood, although it is essential in routine applications, particularly for regulatory compliance.

Both challenges render it difficult to accurately detect low-abundance taxa such as biological contaminations of a sample, especially for complex genera like *Bacillus*. Since the initial characterization of the first *Bacillus* species roughly 150 years ago, the phylogeny of the genus *Bacillus* has reflected continuous advancements in methods deployed for bacterial characterization and identification (Xu & Kovács, 2024). As such, the genus *Bacillus* is still affected by frequent taxonomic shifts, including the promotion of subspecies to species (*e.g.*, *B. subtilis* subsp. *spizizenii* to *B. spizizenii*) (Dunlap et al., 2019), reclassification to other genera (*e.g.*, *B. clausii* to *Shouchella clausii*) (Kim et al., 2023) and addition of new species (*e.g.*, *B. arachidis*) (Chen et al., 2022). For instance, the ongoing reevaluation and reclassification of *B. velezensis* due to genetic and phenotypic resemblance with *B. amyloliquefaciens*, has resulted in many misannotated strains in databases (Kenfaoui et al., 2024; Su et al., 2024; Xu & Kovács, 2024). Yet, in the context of public health, classification to species level is crucial, as only members of the *Bacillus cereus s.l.*, such as *B. anthracis* and *B. cereus sensu stricto (s.s.)*, are pathogenic, while members of the closely related *B. subtilis s.l.*, such as *B. velezensis* and *B. amyloliquefaciens*, are non-pathogenic.

One particular application domain where correct detection of *Bacillus* species is essential, is microbiological fermentation to create commercially available food enzymes (FEs). FEs are enzymes that have a technological purpose at any stage of the manufacturing, processing, preparation, treatment, packaging, transporting, or storing of foods (European Parliament & Council of the European Union, 2008). Microorganisms are the primary producers of enzymes, with the *Bacillus* genus the most prominent source due to its high capacity for protein secretion and diverse traits suited to industrial applications (Danilova & Sharipova, 2020). Genetically modified microorganisms (GMM) can even further enhance the enzyme's purity, yield, specificity, efficiency, stability, and multifunctionality (Zhang et al., 2019). These GMMs often carry antimicrobial resistance (AMR) genes as selection markers to facilitate the genetic modification process. Regulation EC No 1332/2008 on FEs mandates that all FEs must undergo a safety evaluation by the European Food Safety Authority (EFSA) followed by approval from the European Commission prior to authorization for placement on the European market. The regulation guidelines require the applicant to demonstrate the absence of viable cells of the production strain in the final product by means of a culture-based method. In the case of genetically modified (GM) production strains or non-GM strains carrying acquired AMR genes, the absence of DNA should also be demonstrated using PCR with a minimal sensitivity of 10 ng of genomic DNA per gram of product (European Food Safety Authority (EFSA), 2021).

Previous studies have nonetheless highlighted that FEs can be contaminated by (viable) production strains, particularly species from *B. subtilis s.l.* (Deckers et al., 2020; Fraiture, Bogaerts, et al., 2020; Paracchini et al., 2017). Many documented cases of such FE contaminations involve GM production strains that often carry intact AMR genes, posing a potential health risk through horizontal gene transfer. Furthermore, *Bacillus cereus s.l.* is a well-known contaminant in food and feed (Rahnama et al., 2023), and it has been shown that microbial fermentation products can be contaminated with *Bacillus cereus s.l.* (Bogaerts et al., 2023). Due to the similarity in incubation conditions between *B. cereus s.l.* and other *Bacillus* production strains commonly used in microbial fermentation, along with the ubiquity of *B. cereus* and its resilience through spore formation, the microbiological fermentation environment forms a risk for *B. cereus* contaminations that remain in the final product (Jovanovic et al., 2021; Stenfors Arnesen et al., 2008). Because *B. cereus* can cause food poisoning and has been linked to outbreaks and severe clinical manifestation, such contaminations may pose risks to human and animal health as well as economic concerns (Ehling-Schulz et al., 2019). Consequently, detecting potential *Bacillus* contaminants in food enzyme products, whether production strains and/or other contaminants, has become a key area of interest (Fraiture, Deckers, et al., 2020).

The current detection methods for these contaminations are however typically targeted and require prior information and/or microbial isolation. Conventional culture-based methods are still the gold standard but are laborious and time-consuming. Moreover, the microbial contaminants in FE samples are not always culturable due to specific growth needs or auxotrophy. Alternatively, (q)PCR-based methods offer a fast and relatively simple way to amplify and identify specific genetic regions. Deckers et al. developed a conventional PCR which amplifies a region of the 16S-rRNA gene, often used to disseminate the taxonomy of bacterial species, followed by Sanger sequencing for characterization of FE contaminations (Deckers et al., 2020). Although the method detected all tested FE producing bacteria, some such as *B. subtilis* and *B. licheniformis* could only be identified at the genus level. Members of the *Bacillus* genus often carry multiple copies of the same 16S rRNA operon with sequences that frequently overlap across different *Bacillus* species (Strube, 2021). The detection of these multiple variants poses difficulties to distinguish with Sanger sequencing. Moreover, the development of such control methods by enforcement laboratories for post-market control is hampered by the confidentiality of the dossiers submitted to EFSA, and hence the sequences to be looked for are *a priori* not known, rendering their development a laborious and time-consuming case-by-case process. Shotgun metagenomics circumvents these limitations by sequencing all genetic material, providing an open approach of generalized species prediction. Previous studies have demonstrated the added value of metagenomics, showing that it could achieve the same information as the current standard methods while even expanding the characterization of Bacilli producing strains by detecting the presence of additional AMR genes and transgenic constructs (Buytaers, Fraiture, et al., 2021; D'aes et al., 2022). Although these studies demonstrated the potential of metagenomics to detect contaminants in microbial fermentation products, they relied on an extensive in-depth bioinformatics analysis and lacked validated detection thresholds for interpretation, limiting their suitability for routine use. A standardized and validated bioinformatics approach for accurately detecting low-level Bacilli species in FE samples with known LODs and detection thresholds is currently not available.

In this study, we present a general and open bioinformatics framework using metagenomics long-read nanopore sequencing for species classification, applied to *B. subtilis s.l.* production strains and *B. cereus s.l.* contaminants in FE samples. We developed this bioinformatics framework employing *in silico* mixes of isolate nanopore sequencing data, using taxonomic classification for species detection with specific filtering steps and extensive database curation to gain insight into classification performance, to optimize processing steps and to define detection thresholds. The most reliable metric for species detection was template identity, a composite metric that accounts for both the coverage and sequence identity of the matched database entry. The bioinformatics framework was afterwards validated by sequencing five *in vitro* mixes mimicking potential FE contaminations, and its applicability was evaluated by using six real and previously well-characterized contaminated FE samples.

2. Material and methods

2.1. Development of a bioinformatics detection framework using in silico samples

2.1.1. Creation of in silico mixes

2.1.1.1. Generation of isolate nanopore sequencing data. An overview of the workflow for the development of the bioinformatics framework is provided in Fig. 1. Multiple *in silico* mixes were created by subsampling nanopore sequencing reads of five Bacilli species that were cultured and sequenced individually. These five species were chosen based on their occurrence in FE samples as potential contaminations from the production organism(s) (*B. subtilis s.l: B. amyloliquefaciens*, *B. licheniformis*, *B. subtilis* and *B. velezensis*) or as typical food contaminant (*B. cereus s.l.*: *B. paranthracis)*. Three of the species were sourced from the Belgian Coordinated Collections of Micro-organisms, while the remaining two came from an in-house collection, with their corresponding accessions in Table 1.

Each *Bacillus* species was inoculated in 10 ml of Brain Heart Infusion broth (Oxoid, Basingstoke, UK) and incubated at 35 °C for 24 h. Bacterial cells were collected by centrifugation (5 min at 6000 xg) and DNA was extracted as described before (Gand et al., 2024), with 2 h of metapolyzyme (Sigma–Aldrich, Saint-Louis, MA, USA) incubation. DNA quality and quantity was evaluated using the NanoDrop 2000 (Thermo Fisher Scientific, Waltham, MA, USA) and Qubit 4.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA), respectively.

Library preparation was performed using the Ligation sequencing DNA kit (SQK-LSK110, Oxford Nanopore Technologies, Oxford, UK) following the manufacturing instructions. Each sample was sequenced on a separate ONT flow cell R9.4.1 (FLO-MIN106) in a GridION Mk1 device for 72 h. to ensure high sequencing depth was available for creating the *in silico* mixes. The isolate sequences were basecalled with Dorado 0.7.0 with model *dna_r9.4.1_e8_sup@v3.6* (Oxford Nanopore PLC, 2023). Quality metrics of the sequences for each isolate are shown in Table 1.

2.1.1.2. In silico length transformation. Due to the downstream processing of FE products leading to DNA degradation, both the sequencing quality and lengths of the isolates were compared against five metagenomic sequenced FE products (D'aes et al., 2022). In a prior study, these commercial FE products were sequenced with R9 Nanopore to conduct a metagenomic characterization of GM *Bacillus* contaminations (D'aes et al., 2022). The read statistics for the FE products along with their accession numbers are provided in Supplementary Table S1. While the quality distribution between the isolates and metagenomic FE products was similar, the length distribution differed substantially (Supplementary Fig. S1). The sequences from the metagenomic FE products were shorter than those from the isolates. As read lengths influence classification performance (Portik et al., 2022), the isolate sequencing reads were not randomly subsampled. Instead, they were subsampled based on the read length to match their read length distributions with those of real metagenomic FE distributions, ensuring that the *in silico* mixes closely resembled actual metagenomic read length distributions.

The subsampling strategy involved three steps (Supplementary Fig. S2). First, all sequencing reads from the metagenomic samples were used without prior quality trimming and grouped into length-based bins with a bin width of 100. The reads from an individual isolate were also similarly grouped into bins. Each metagenomic bin was then assigned a weight proportional to the number of reads it contained. Second, rather than randomly subsampling from the isolate, the metagenomic bins were subsampled according to their respective weights. Then, for each selected metagenomic bin, a random read from the corresponding isolate bin was chosen. If the bin number was 10 or less, the read was selected with replacement because reads shorter than 1000 will be filtered out later on. If the bin number exceeded 10, it was selected without replacement. In cases where the corresponding isolate bin lacked reads or was already empty due to previous selections, the bin number was retained for the next step. Third, the remaining bin numbers that lacked a matching sequence from the isolate bins were arranged in descending order. For each of these remaining bins, if the bin number was higher than any non-empty isolate bin, a random sequence was selected from the closest lower isolate bin without replacement. If no lower bin was available, a random sequence was selected from the closest higher isolate bin without replacement and randomly clipped to

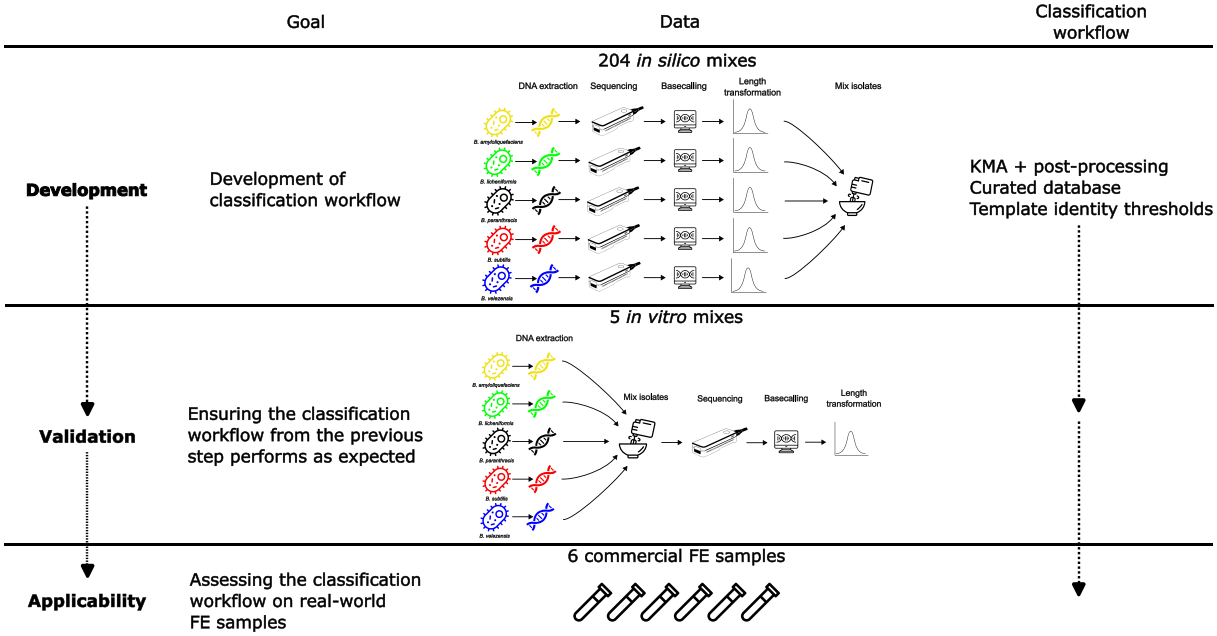

Fig. 1. General overview of the workflow. For the development step, the bioinformatics framework was developed using 204 *in silico* mixes. Taxonomic classification was done using KMA and several post-processing steps, including filtering on template identity and using a curated reference sequence database. Next, the bioinformatics framework was validated using five *in vitro* mixes. Lastly, the applicability of the bioinformatics framework was evaluated on six commercial FE samples.

Table 1.
Sequencing metrics of the five *Bacilli* isolates used for creating the *in silico* mixes. The first and second column denote the species group and species of the sequenced isolate. The third column contains either the BCCM accession or an in-house Sciensano accession of the isolate. The fourth and fifth column represent the total number of reads and bases, respectively. The, sixth, seventh and eighth column show the average length, median length and N50, respectively. The final column presents the average Phred score.

Species group	Species	Accession number	Number of reads	Number of bases	Average length	Median length	N50	Average Phred score
B. subtilis s.l.	*B. amyloliquefaciens*	LMG 12325	7,777,345	39,311,153,315	5054.6	2083	13,188	11.15
	B. licheniformis	LMG 7558	1,307,714	16,491,117,680	12,610.6	7031	26,611	11.32
	B. subtilis	Sciensano E07–505	2,314,142	6,797,983,369	2937.6	1385	6122	12.08
	B. velezensis	LMG 22478	3,319,078	31,252,767,934	9416.1	5061	20,299	11.64
B. cereus s.l.	*B. paranthracis*	Sciensano 41	16,966,103	25,867,944,882	1524.7	661	3260	11.11

Abbreviations: BCCM (Belgian Coordinated Collections of Micro-organisms).

fit the appropriate length for the bin.

2.1.1.3. Composition of the 204 in silico mixes for developing the detection strategy. A total of 204 *in silico* mixes were generated from the isolates, each with a yield of 3 million (M) reads (corresponding to roughly 2250 M bases), in concordance with the expected yield of the five real metagenomic FE samples (Supplementary Table S1). These mixes were categorized into three types (Supplementary Fig. S3). The first type included two species, each comprising 50 %. The second type included three species: the first two species, which could not be *B. paranthracis*, each contributed 45 % of the mix, while *B. paranthracis* made up the remaining 10 % as the third species. The third type consisted of two species, where the first species constituted between 90 % and 99 % in steps of 1 %, and the second species completed the remaining percentage up to 100 %. The three types resulted in 6, 6 and 192 mixes, respectively. A full overview of all permutations is also available in Supplementary Table S2. For each mix, every isolate was subsampled to the specified abundance as described in Section 2.1.1.2 and then combined together (Fig. 1).

2.1.2. Taxonomic classification

Species were detected with a taxonomic classifier, which assigns a taxonomic identification to each read by comparing it against a reference database containing representative sequences. A good classifier should identify all species that are actually present, producing true positives (TPs), while minimizing the species incorrectly identified as present, referred to as false positives (FPs). A truly present species that goes undetected is considered as a false negative (FN). To reduce FPs, thresholds are typically applied to decide whether a detected species should be considered present. However, applying these thresholds can also lower TPs, turning them into FN if they fall below the threshold. The impact of a threshold is typically evaluated using precision and recall. Precision is the proportion of detected species that are truly present and improves as FPs decrease, while recall is the proportion of truly present species that are correctly identified and improves when FNs decrease. Together, these metrics highlight the trade-off between detecting all present species and avoiding incorrectly detected species. Based on previous work, KMA was selected as taxonomic classifier for nanopore reads for DNA-to-DNA classifications for this study because it offers very high recall with the lowest penalty in precision (Clausen et al., 2018; Van Uffelen et al., 2024).

Before classification, sequencing reads were filtered with SeqKit 2.3.1 on a length and quality higher than 1000 and 7, respectively (Shen et al., 2016). Classification was performed with KMA 1.4.12a using the parameters "-mrs 0.0", "-bcNano", "-bc 0.7", "-ef", "a", "-mem_mode", "-1t1", "-matrix", "-tsv" and "-shm". After classification, three additional postprocessing steps were applied. First, all reads classified as a plasmid were removed. Due to the mobility of plasmids, they offer low taxonomic discriminatory power and can distort classification results. The method of determining whether a reference sequence was a plasmid depended on the sequence's origin (see Section 2.1.3). For reference sequences from NCBI RefSeq, a sequence was considered a plasmid if its name contained the word "plasmid". For sequences from BTyperDB (see Section 2.1.3), two scenarios were considered, as BTyperDB utilizes both publicly available assemblies from NCBI and its own assemblies. For NCBI assemblies, sequence names were not used for plasmid identification due to GenBank's less stringent quality standards compared to RefSeq. Instead, the field "assigned_molecule_location_type" from NCBI's API "sequence_reports" was used to identify plasmids (O'Leary et al., 2024). For BTyperDB's own assemblies, no plasmids were identified because annotation information was unavailable. Second, reference sequences from the same genome were combined and the reported metrics were recalculated for the entire genome rather than for each sequence individually. A sequence that initially appears to be a good match may not hold up when evaluated against the entire genome, since genomes in the database can consist of multiple sequences. Per reported reference genome, the metrics were recalculated by the arithmetic mean weighted by the sequence lengths:

$$Recalculated\ metric = \frac{\sum_{i=1}^{n} L_i x_i}{\sum_{i=1}^{n} L_i} \quad (1)$$

with L the sequence length, x the metric, i a specific sequence of the genome and n the number of sequences in the genome. Third, for each species, only the genome with the highest chosen metric was retained, while the others were discarded. Amongst the possible metrics for distinguishing presence from absence, template identity was the most effective (Supplementary Text S1.1 and Supplementary Fig. S4). The template identity is defined as:

$$Template\ identity = \frac{\#Identical\ Bases(Consensus, Template)}{Template\ Length} \quad (2)$$

with the consensus reflecting the majority vote from all aligned reads and the template being the reference in the database. Template identity differs from the traditional definition of identity by considering not just identity over the aligned length but rather the entire length of the template.

2.1.3. Reference database

Assemblies from the NCBI Reference Sequence Database (RefSeq) were used as a database for the classification (accessed on 24/02/2023) (O'Leary et al., 2016). The selected assemblies included the taxonomic groups Archaea, Bacteria, Fungi, Protozoa and Viruses. Assemblies for Bacteria and Viruses were required to be at the 'complete' level, whereas assemblies for the other groups needed to be of a level higher than or equal to 'scaffold'. Additionally, the human genome was added. Afterwards, three filter steps were applied to increase the quality of the database. First, all assemblies from any taxonomic groups that included the word 'unclassified' were removed. Assemblies in such taxonomic groups are often of substandard quality and their true species assignment is unknown. Second, all assemblies from the *Bacillus cereus* group

(*sensu lato*) were replaced with assemblies and their species annotations from BTyperDB (Ramnath et al., 2023). The *Bacillus cereus* group is a taxonomic complex group for which genomes on NCBI are not always correctly labeled (Carroll et al., 2022). BTyperDB is an atlas of *B. cereus s.l.* genomes with standardized, community-curated metadata with many novel and previously unassembled genomes. Replacement of the *B. cereus s.l.* genomes from NCBI with those from BTyperDB was hence done to improve the quality of this species group. However, since BTyperDB uses a revised taxonomic nomenclature for *Bacillus cereus s.l.*, their final taxonomic assignments could not be used directly in our workflow, which relies on the NCBI taxonomic nomenclature. Therefore, we assigned species-level taxonomy to these genomes based on BTyperDB's calculated average nucleotide identity (ANI) against type strain genomes of NCBI's *B. cereus s.l.* species, adhering to NCBI's taxonomic framework. Third, the assemblies from the *Bacillus subtilis* group (*sensu lato*) were curated based on ANI values and hierarchical clustering. Like the *B. cereus* group, the *B. subtilis* group suffers from incorrectly labeled genomes in NCBI (Chorlton, 2024; Xu & Kovács, 2024). Because there is no equivalent of BTyperDB for *B. subtilis s.l.,* a custom in-house curation approach was taken. The ANI values between all *B. subtilis s.l.* RefSeq genomes were calculated with fastANI 1.33 using default parameters (Jain et al., 2018). The resulting ANI values were then hierarchically clustered in Python 3.10 using the 'clustermap' function from Seaborn v0.13 with default settings (Waskom, 2021). Clusters were demarcated by an ANI threshold higher than 98 % as shown in Supplementary Fig. S5. The 37 resulting clusters were each assigned a species based on majority vote. Genomes within clusters that did not match the majority species were discarded. As a result, 140 *B. amyloliquefaciens* genomes across two clusters dominated by *B. velezensis* and two *B. velezensis* genomes across two clusters dominated by *B. amyloliquefaciens* were removed. The three curation steps resulted in a database with 43,456 genomes containing 3,285,670 sequences, of which 6648 genomes belonged to 59 different *Bacillus* species (see Supplementary Table S3).

2.2. Validation of the developed bioinformatics detection framework using in vitro samples

2.2.1. Creation of in vitro mixes

2.2.1.1. Generation of in vitro metagenomic nanopore sequencing data. To validate the developed bioinformatics framework, five *in vitro* mixes were created. For the *in vitro* mixes, the DNA of the five isolates (see Section 2.1.1.1) was separately extracted, mixed together and sequenced. The full workflow of the creation of the *in vitro* mixes is depicted in Fig. 1 under the validation step. Table 2 outlines the composition of the five *in vitro* mixes, along with the motivation behind each specific mixture. Supplementary Table S4 contains an overview of the DNA amounts that were mixed. Library preparation was performed using the Ligation sequencing DNA V14 kit (SQK-LSK114, Oxford Nanopore Technologies, Oxford, UK) following the manufacturing instructions. Each *in vitro* mix was sequenced on a separate ONT flow cell R10.4.1 (FLO-MIN114) in a GridION device for 72 h. Each *in vitro* mix was basecalled with Dorado 0.7.0 with the model *dna_r10.4.1_e8.2_400bps_sup@v4.3.0*. Table 3 outlines the sequence metrics of each *in vitro* mix.

2.2.1.2. In vitro length transformation. Although the *in vitro* mixes underwent metagenomic sequencing, their read lengths differed from those of real metagenomic samples due to additional DNA degradation factors that cannot be accurately replicated *in vitro* (Supplementary Fig. S1). However, applying the same length transformation procedure used for the 204 *in silico* mixes to the *in vitro* mixes was not possible. For *in silico* mixes, the transformation could be applied to the sequencing reads of each isolate individually before combining the transformed reads to create the final mix, preserving the appropriate relative species abundances within the mixtures. Applying the length transformation after mixing would have distorted the relative species abundances. Similarly, directly applying the read length transformation to the five *in vitro* mixes would alter species abundances. Therefore, an alternative strategy was employed to correct the length distribution of the *vitro* mixes by transforming the reads of each species individually based on the taxonomic allocation of each read.

Table 2.
Composition of the five *in vitro* mixes. The first column denotes the mix number. The second and third column show the species and their relative abundance in a mix. The last column provides a concise motivation behind the construction of a mix.

Mix	Species	Relative abundance (%)	Motivation
1	*B. amyloliquefaciens* *B. licheniformis* *B. velezensis*	75 22.5 2.5	Similar composition observed in practice (D'aes et al., 2022). Contains *B. velezensis* as low-level contaminant at 2.5 %.
2	*B. licheniformis* *B. velezensis*	80 20	Similar composition observed in practice (D'aes et al., 2022).
3	*B. amyloliquefaciens* *B. licheniformis* *B. subtilis*	76.5 22.5 1	Copy of mix 1 where *B. amyloliquefaciens* and *B. licheniformis* are reversed as majority species. Also contains *B. subtilis* as low-level contaminant at 1 %.
4	*B. amyloliquefaciens* *B. licheniformis* *B. paranthracis*	75 20 5	Copy of mix 1 with *B. paranthracis* as low-level contaminant at 5 %.
5	*B. amyloliquefaciens* *B. licheniformis* *B. paranthracis* *B. subtilis* *B. velezensis*	20 20 20 20 20	All species from *B. subtilis s.l.* and *B. paranthracis*.

The taxonomic labels of the reads in the *in vitro* mixes were determined by mapping the reads to the closest reference genomes of the species present in the mix. To determine the closest reference genomes for each species, the Mash distances between the full yield of each isolate (see Section 2.1.1.1) and all complete RefSeq assemblies under the genus *Bacillus* were calculated using Mash 2.3 with parameters "-r", "-m 2" and "-s 10000" (Ondov et al., 2016). For each species, the genome with the most shared hashes (*i.e.*, lowest distance) was selected as reference genome. Prior to mapping the reads of the *in vitro* mixes, reads shorter than 15,000 bases were filtered out to minimize the risk of incorrect mappings. The remaining reads from each *in vitro* mix were subsequently mapped to an index of the reference genomes for the species within the mix using Minimap v2.26 with the parameter '-x map-ont' (Li, 2018). To further reduce the risk of incorrect mappings, four postprocessing steps were implemented. First, reads that mapped to multiple reference genomes were removed. Second, reads with at least one mapping with a MAPQ score below 60 were filtered out. Third, reads with overlapping mapping sections were discarded. Fourth, reads with less than 80 % coverage were removed. Supplementary Table S5 provides details on the number of bases filtered out at each step, as well as the remaining number of bases after all steps. Supplementary Table S6 shows the number of reads and bases mapped to each species' reference genome per mix after the processing steps. A comparison of the species' breadth of coverage after mapping all unfiltered reads *versus* filtered reads revealed that no genomic regions were lost during the filtering process (Supplementary Table S7).

Table 3.
: Read metrics of the different types of *in vitro* mixes and corresponding *in silico* mixes. The first and second column denote the mix type and number with the composition shown in Table 2. The third and fourth column show the number of reads and bases. The fifth, sixth and seventh column show the average length, median length and N50 of the corresponding mix. The last column displays the average Phred read score.

Type of mix	Mix number	Number of reads	Number of bases	Average length	Median length	N50	Average Phred score
In vitro (raw)	Mix 1	5,142,921	32,668,427,519	6352.1	1796	22,009	13.67
	Mix 2	2,422,191	30,979,842,193	12,790	7524	26,147	13.81
	Mix 3	5,661,184	24,974,818,719	4411.6	1564	14,811	13.74
	Mix 4	3,802,777	24,644,907,350	6480.8	1615	24,003	13.04
	Mix 5	11,075,295	28,588,330,396	2581.3	946	7576	13.39
In vitro (length transformation)	Mix 1	3,000,661	2,240,033,700	746.5	533	958	14.09
	Mix 2	2,997,815	2,241,347,200	747.7	533	959	14.16
	Mix 3	2,995,740	2,240,059,505	747.7	532	959	14.31
	Mix 4	2,994,299	2,236,673,892	747	532	959	13.75
	Mix 5	2,997,281	2,238,044,518	746.7	533	957	14.27
In vitro (no length transformation)	Mix 1	352,564	2250,000,493	6381.8	1806	22,139	13.65
	Mix 2	176,094	2250,012,072	12,777.3	7517	26,101	13.82
	Mix 3	509,235	2250,004,478	4418.4	1564	14,880	13.68
	Mix 4	346,714	2250,006,084	6489.5	1618	24,041	13.06
	Mix 5	867,774	2250,024,991	2592.9	946	7657	13.39

The taxonomically labeled reads from the *in vitro* mixes were then used to create length transformed *in vitro* mixes with the same abundances as described earlier in Table 2 and with the same yield as the 204 *in silico* mixes. The subsampling approach used for the length transformation of the *in silico* mixes (see Section 2.1.1.2) could not be applied here, as some species did not have enough reads. However, all species had sufficient sequencing yield in total amount of bases, allowing larger reads to be split into multiple smaller reads. Subsequently, a slightly modified version of the length transformation was applied. The first and second steps followed the same approach. Reads from bin numbers lower than 10 could never be selected with replacement, as all reads from the species were all above 15,000. The main difference occurred in the final part of the third step. When no lower bin was available in the species' bins, a random sequence was not selected without replacement but instead clipped to the appropriate length and reassigned to the correct bin based on its updated length. Additionally, because the reads were split, the final yield of the transformed *in vitro* mixes needed to be expressed in number of bases rather than number of reads. On average, the *in silico* mixes with 3 M reads contained 2250 M bases, meaning the final yield of the *in vitro* mixes was set to 2250 million bases.

2.2.2. Validation of the developed bioinformatics workflow

Two types of subsampled *in vitro* mixes were prepared: one with length transformation (see Section 2.2.1.2) and one without, both with the same yield as the 204 *in silico* mixes, for which the read statistics can be found in Table 3. Comparison of both datasets was performed to evaluate if longer read lengths (*i.e.*, without length transformation and not representative for real FE samples) indeed resulted in a higher performance. On both datasets, the developed bioinformatics framework was run, using the same taxonomic classification with KMA and curated reference database, followed by post-processing and template identity filtering.

2.3. Application of the developed and validation bioinformatics detection framework using real samples

Lastly, the applicability of the developed and validated method was also evaluated using real FE samples, depicted in Fig. 1 under the applicability step. Six FE samples from a previous study were processed with the developed strategy (D'aes et al., 2025). The original study sequenced these samples with an R9 flow cell and basecalled with Guppy 5.0.7 in GPU mode, using the *dna_r9.4.1_450bps_sup* model. Species detection in the original study was conducted through a comprehensive in-depth analysis that integrated multiple methods, including metagenomic hybrid assembly. This approach produced a high-quality set of species labels for each sample, serving as the ground truth for evaluating our open detection strategy in our study. A full overview of the samples, including their accession numbers and original read statistics, is available in Supplementary Table S1. Afterwards, the samples were randomly downsampled to a yield of 3 M reads, except for two samples that already had fewer than 3 M reads (2.6 M and 2.3 M), and then processed with the developed and validated bioinformatics workflow described above.

3. Results

Fig. 1 provides an overview of the workflow of our study, divided into three steps: development, validation and applicability evaluation of a bioinformatics framework for the detection of sample contaminants, applied to *B. subtilis s.l.* production strains and *B. cereus s.l.* contaminants in FE samples. For the development, 204 *in silico* mixes were used to gain insight in classification performance, fine-tune processing steps and define general detection thresholds. The developed classification framework was afterwards applied to 5 *in vitro* mixes to validate its performance. Lastly, the applicability was evaluated using six commercial FE samples as real-world examples.

3.1. Development of an open bioinformatics detection framework

To develop the detection framework, an *in silico* mixture modeling approach was employed using nanopore sequencing data (Fig. 1). A total of 204 *in silico* mixes (see Supplementary Fig. S3) were created based on five different Bacilli reference samples that underwent isolate nanopore sequencing (Table 1). Before mixing the isolate reads *in silico,* they first underwent a length transformation to account for the shorter nanopore read lengths observed in real metagenomic FE data whilst also mimicking sequencing yields observed in practice (Supplementary Fig. S1 and Supplementary Fig. S6). The 204 *in silico* mixes were subsequently used to develop the detection strategy, which consisted of KMA complemented with additional filtering and post-processing steps whilst using a custom reference sequence database that was curated to contain only high-quality Bacilli genomes.

Fig. 2 shows a parallel plot with key metrics for the species identified after KMA classification of all 204 *in silico* mixes. Notably, no FNs were detected in any of the *in silico* mixes, even for species present in the mixes at 1 %. However, as expected, all mixes except for one exhibited at least one FP, being either of the genus *Bacillus* or a *Bacillus* phage. Of the different metrics evaluated, neither number of aligned reads, nor depth,

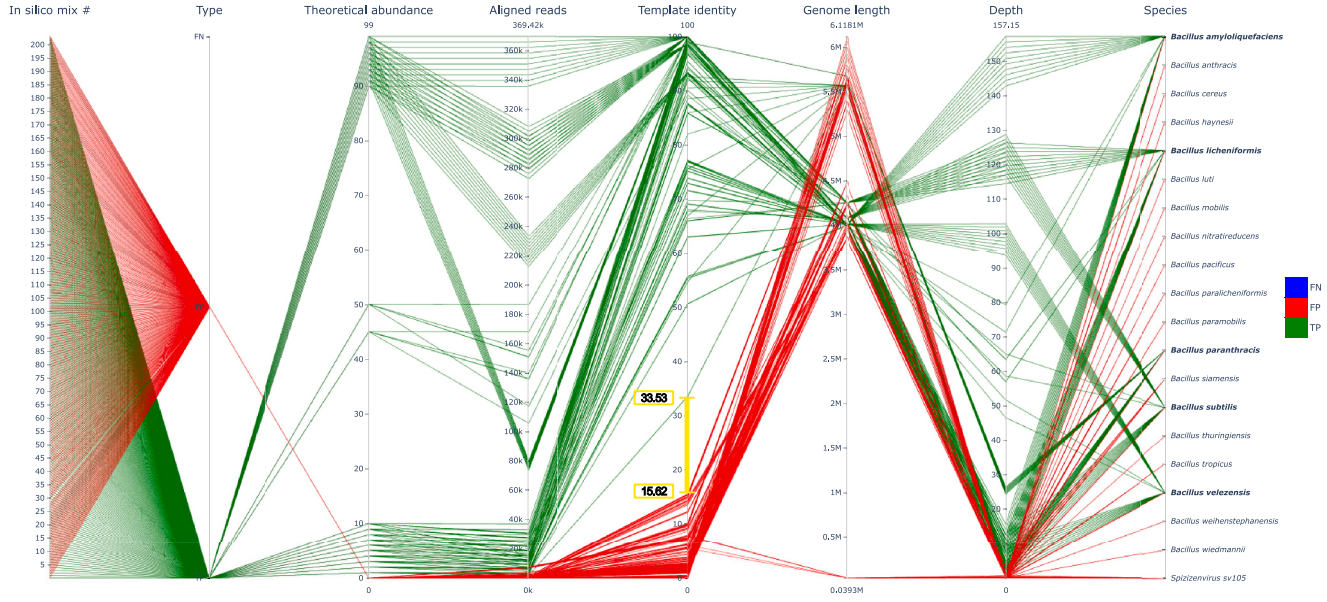

Fig. 2. Parallel plot of the results of taxonomic classification with KMA for all 204 *in silico* mixes during development of the bioinformatics framework. Each vertical line represents a relevant metric associated with a reference genome detected by KMA, while each horizontal line corresponds to a classification result. The intersection of the horizontal line with the vertical lines provides the value of a given metric for a given *in silico* mix. Colors represent the type of classification result: green for TP, red for FP and blue for FN. The first vertical line indicates the *in silico* mix. The second line specifies the type of hit: TP, FP or FN. The third line shows the relative theoretical abundance based on the ground truth (note that all FP were allocated a theoretical abundance of 0 %). The fourth, fifth and seventh column represent the number of aligned reads, template identity and depth as calculated by KMA, respectively. The sixth and eight lines display the length and species of the reference genome. Species that were included in the *in silico* mixes are indicated in bold. The yellow line marks the range between the lowest and highest template identities for the TPs and FPs, respectively, with both values indicated. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
Abbreviations: FP (false positive), TP (true positive), FN (false negative).

nor genome length, allowed a strict separation of TPs from FPs (Supplementary Fig. S4). However, template identity effectively allowed clear separation of TPs from FPs in all 204 *in silico* mixes. The highest template identity of any FP in all *in silico* mixes was 15.62 % (mix with 99 % *B. paranthracis* and 1 % *B. subtilis*) while the lowest template identity of any TP was 33.53 % (mix with 1 % *B. paranthracis* and 99 % *B. subtilis*). As such, template identity was integrated into the detection strategy as a metric for detecting Bacilli, with a range of 17.91 % (between 15.62 % and 33.53 %) that effectively distinguished FPs from TPs. The results indicated that the LOD based on the *in silico* findings was at least 1 % for these species mixtures (lower than 1 % was not evaluated).

Given the feasibility of employing template identity for separating TPs from FPs in the *in silico* mixes, the potential confounding effects of individual species relative abundances and interaction effects between different species were further investigated. Fig. 3A shows the median

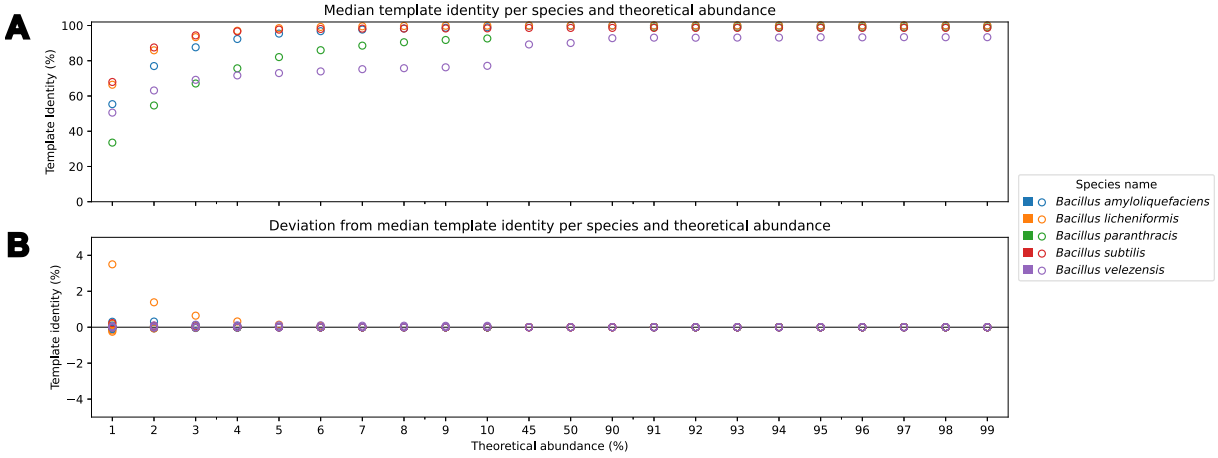

Fig. 3. A) Median template identity per species and theoretical abundance. For each theoretical abundance on the x-axis, the y-axis shows the median template identity for the species. B) Deviation of median template identity per species and theoretical abundance. For each theoretical abundance on the x-axis, the y-axis shows the deviation of median template identity relative to the median template identity of the corresponding species (shown in plot A). Points with a positive deviation lie above the horizontal black line while points with a negative deviation lie below the horizontal line.

template identity for each species across its different theoretical abundances. With equal yield, the template identity remained stable and showed only minor differences at high abundances. However, at lower abundances, the difference in template identity became more pronounced with some species showing a more noticeable decrease than others. Fig. 3B shows the deviation in template identity for each species at different theoretical abundances, compared to their median template identity at the same abundance level. The template identity of the same species with the same relative abundance remained consistent, regardless of species composition. However, at lower abundances, deviations in template identity for the same species became more pronounced due to small fractions of erroneous mappings (results not shown).

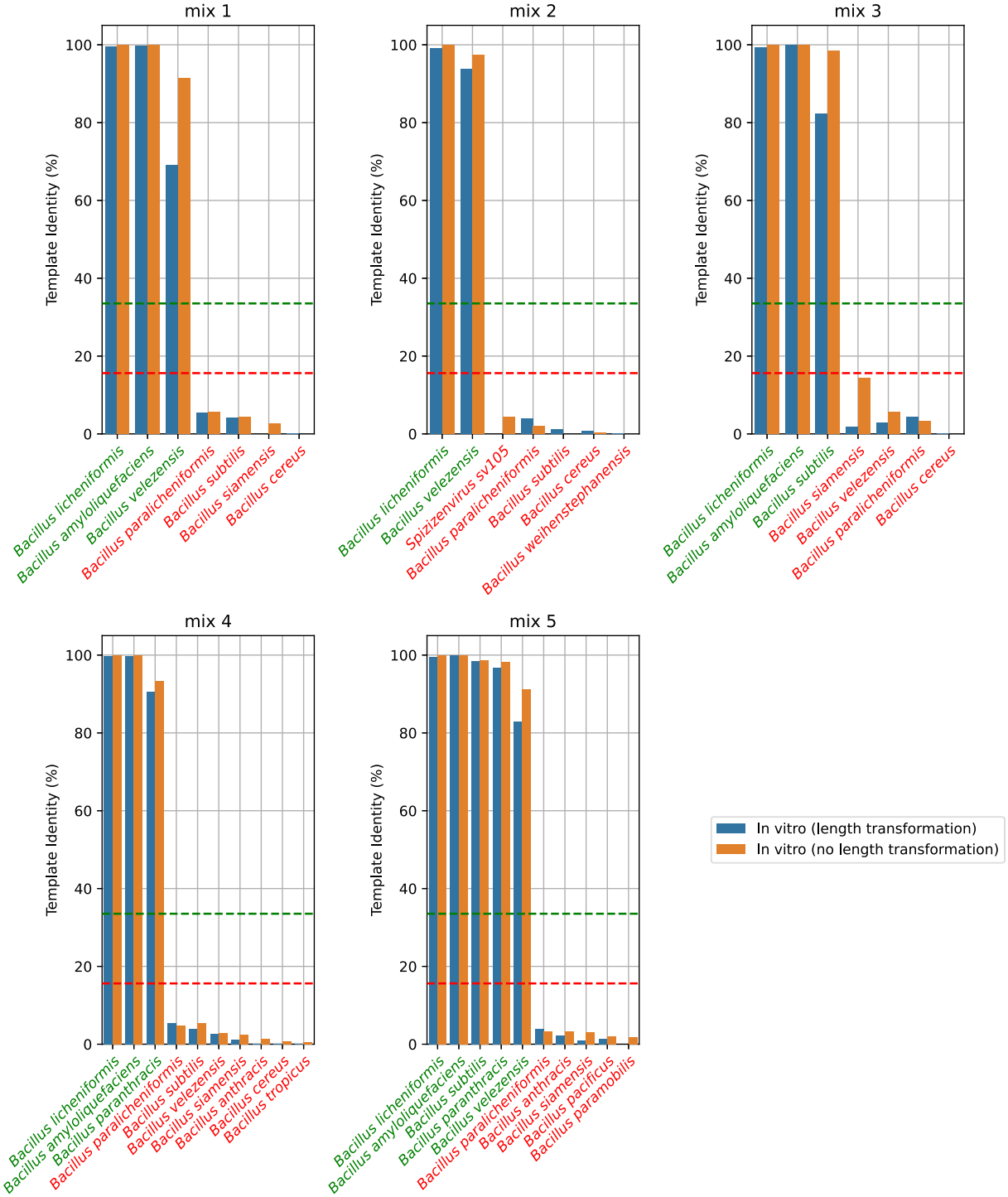

Fig. 4. Template identity for the five *in vitro* mixes with and without length transformation. Each panel represents an *in vitro* mix (Table 2) with and without length transformation, both normalized to the same yield. The x-axis represents the top 10 highest scoring species (or fewer if less were detected) marked green for TPs and red for FPs. The y-axis depicts the template identity (%). The red and green dotted lines represent the template identity interval between 15.62 % and 33.53 % template identity, below which only FPs are detected and above which only TPs are detected. No FN were detected in any of the *in vitro* samples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
Abbreviations: FP (false positive), TP (true positive), FN (false negative).

Consequently, using the tool KMA for taxonomic classification with specific post-filtering steps (see Section 2.1.2), coupled with the in-house curated reference sequence database (see Section 2.1.3), was retained as bioinformatics framework for species detection. In particular, species detected with a template identity lower than 15.62 % were dismissed as FP, and species detected with a template identity higher than 33.53 % were retained as TP. Species detections with a template identity between these values merit more in-depth *ad hoc* bioinformatics investigation. At a yield of 3 M reads, the LOD for accurate detection was 1 %.

3.2. Validation of the developed bioinformatics detection framework

To validate the developed bioinformatics detection framework and associated template identity thresholds, individually extracted DNA of each target species was mixed before undergoing direct metagenomic nanopore sequencing (Fig. 1) for a total of five *in vitro* mixes (Table 1). To account for the impact of sequencing yield on classification results (see Supplementary Fig. S7), these five *in vitro* mixes were downsampled to a similar yield as employed in the *in silico* mixes. Additionally, as their read length distributions, similar to the isolate sequencing data, displayed higher read length distributions compared to real FE samples (see Supplementary Fig. S1), they were also length transformed. Since the length transformation process could potentially induce a bias in this case (since reads in the *in vitro* mixes had to be first sorted bioinformatically before read length transformation was possible, in contrast to the *in silico* mixes), a dataset not employing read length transformation was also retained. Both the *in vitro* mixes with and without read length transformation were subsequently classified with KMA in the same manner as described in the previous section to evaluate whether species could be correctly detected using the established range of [15.62 %, 33.53 %] for template identity. Fig. 4 shows the detected species in each mix, with and without length transformation. In general, all TPs were identified and no FNs were observed using template identity. Notably, the template identity of the mixes with length transformation was slightly lower than those without length transformation. The lower scores of the length-transformed mixes confirmed the *in vitro* length transformation process did not induce a bias by selecting high-quality, long-read mappings during the transformation and moreover emphasized the added value of the length transformation. The transformed data contained a higher proportion of shorter reads that rendered classification more challenging, as is expected in real shotgun metagenomics data. Consequently, for the remainder of the validation, only the length-transformed mixes will be discussed.

In all *in vitro* mixes, several FPs were detected but always at low template identity values, therefore allowing again for a clear separation between TPs and FPs. The ranges of template identities to separate FPs and TPs were [5.34 %, 69.08 %], [3.89 %, 93.71 %], [4.38 %, 82.24 %], [5.45 %, 90.61 %] and [3.95 %, 82.84 %] for mixes one to five, respectively. The lower template identities for *B. velezensis* (69.08 %) and *B. subtilis* (82.24 %) in mix 1 and mix 3, could likely be partly attributed to their lower abundances of 2.5 % and 1 %, respectively. This trend was also apparent in Fig. 3A, which showed a consistently lower template identity at low abundances. However, the lower template identity of *B. velezensis* (82.84 %) in mix 5 could not be explained by its abundance alone, which was 20 %. Instead, it likely reflected the generally lower template identity of *B. velezensis* at any abundance compared to the other Bacilli, also depicted in Fig. 3A.

As such, the previously developed bioinformatics framework using specific filtering steps and the in-house curated database was validated successfully using the five *in vitro* mixes. The associated established *in silico* template identity spectrum for distinguishing TPs from FPs of [15.62 %, 33.53 %] by considering species detections above and below the template identity thresholds of 33.53 % and 15.62 % as present and absent, respectively, was also validated with the five *in vitro* mixes.

3.3. Applicability of the open bioinformatics detection framework using real samples

To test the applicability of the developed and validated bioinformatics framework, including the employed template identities for rejecting species detections as FP or accepting them as TP, six commercial FE samples were used, fully characterized in a previous study (D'aes et al., 2025). Species detection in the original study involved an exhaustive in-depth analysis using metagenomics hybrid assembly and additional strain-level analysis, providing a high-quality set of species labels as ground truth for evaluating our open detection strategy. Fig. 5 shows a heatmap with the template identity of the detected Bacilli species per sample (see Supplementary Table S8 for all results). Detected Bacilli with a template identity below the earlier defined threshold of 15.62 % were considered absent and are therefore not depicted, unless the species was part of the ground truth (*i.e.*, a FN). Hatched cells indicate ground truth species, with their relative abundance as found by D'aes et al. between brackets. Employing our framework resulted in correct detection of all expected Bacilli for four out of six samples (M1, A12, P1 and A4). In the other two samples (A2 and A3), one expected species was observed below the 15.62 % threshold, while two others were found within the twilight zone between [15.62, 33.53]%. In sample A2, *B. velezensis* had a template identity of 3.84 %. This species was originally found to be present by D'aes et al., but at a very low relative abundance of only 0.5 %, hence below the LOD of 1 % our framework. *B. amyloliquefaciens* had a template identity of 21.5 %, below the threshold of 33.53 % for reliable species detection but above the threshold of 15.62 % for species absence in our framework. This species was however originally found to be present also at a low relative abundance of 1.5 % by D'aes et al., close to the LOD of 1 % of our framework. The same was observed in sample A3, where *B. velezensis* had a template identity of 33.3 %, just below the threshold of 33.53 % for reliable species detection, but this species was originally found to be present at a low relative abundance of 1.4 %, again close to the LOD of 1 % of our framework. In particular, both samples also showed lower yields after filtering compared to the in silico mixes (Supplementary Fig. S6), explaining the difficulty in species detection near the 1 % LOD. Notably, *B. velezensis* was reliably detected in sample A4, which has a similar filtered yield and length distribution as A3, despite its abundance of 1.7 % close to the LOD. This demonstrates that even a small increase in low-abundance species can have a large impact on the template identity, as shown in Fig. 3A, and that a minimum sequencing yield is required to meet the 1 % LOD of our framework (i.e., 3 M reads or 950 M bases). Although no FPs were detected, two samples (A3 and A4) unexpectedly contained *B. paralicheniformis* in the ambiguous range between 15.62 % and 33.53 %. While *B.* paralicheniformis was never detected with a template ID above 6.25 % in the in silico mixes, its close relatedness to *B. licheniformis* (Du et al., 2019; EFSA Panel on Food Contact Materials, Enzymes and Processing Aids (CEP) et al., 2024), combined with both the high abundance of *B. licheniformis* and the smaller read lengths (See Supplementary Table S9 and Supplementary Fig. S6) within the two samples, led to overestimated template identity for *B. paralicheniformis*. This was further supported by the low alignment scores of the reads mapped to *B. paralicheniformis* by KMA (See Supplementary Fig. S8).

In conclusion, the applicability analysis confirmed that our open workflow is capable of accurately detecting *B. subtilis* s.l. and *B. cereus* contaminants in real FE samples up to an abundance of 1 %, provided a post-filtering sequencing yield of at least 950 M bases is available. Species present with a template identity in the interval [15.62 %–33.53 %] are recommended to be investigated with additional ad hoc bioinformatics analysis or qPCR to investigate whether these are potentially present at very low abundances.

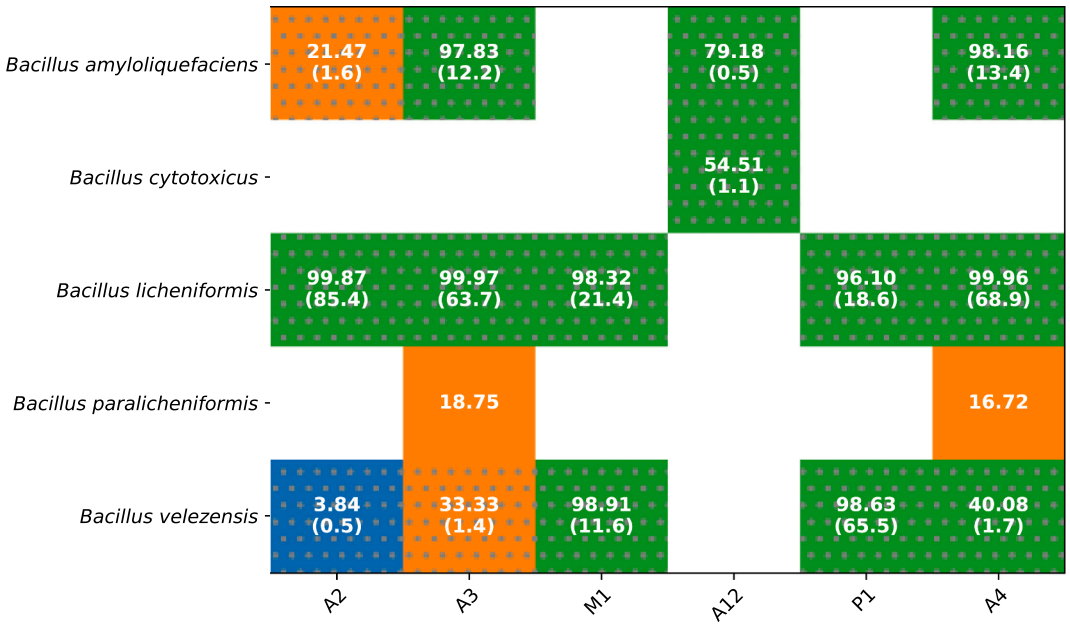

Fig. 5. Heatmap of the detected relevant Bacilli species in real FE samples. Hatched cells constitute the ground truth species present in samples with their relative abundance between brackets, as described in the original study (D'aes et al., 2025). The values represent the template identity of the Bacilli as detected by the developed and validated bioinformatics framework of this study. Green and blue cells represent TPs and FP, respectively, while orange cells warrant further investigation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
Abbreviations: FP (false positive), TP (true positive), FN (false negative).

4. Discussion

In this study, we developed and evaluated an open bioinformatics framework to detect low-level biological contaminations in samples using nanopore shotgun metagenomic data, applied to *B. subtilis s.l. and B. cereus* contaminants in FE samples. The implementation of the framework uses the tool KMA for taxonomic classification with specific post-filtering steps and a curated database of Bacilli genomes. Template identity could be used to accurately predict TP species up to a LOD of 1 % without introducing FP species when a minimum post-filtering yield of 950 M bases was available. Real-world applicability was confirmed using commercial samples.

The read length distribution of the *in silico* mixes was length transformed, as the read length of metagenomic FE samples is typically substantially shorter than isolate sequencing (Supplementary Fig. S1), and shorter read lengths penalize classification performance (Govender & Eyre, 2022; Portik et al., 2022). Template identity proved to be the most effective metric to separate TPs from FPs without introducing FNs (Fig. 2, Supplementary Fig. S4). Species with a template identity above 33.53 % were considered as present, and those below 15.62 % considered as absent. The effectiveness of template identity primarily stems from the inherent verification of the accuracy of read alignments (see Eq. (1)). Template identity can be considered a stricter measure than the breadth of coverage, as it also accounts for the accuracy of the paired bases. In a mix where a TP has a very high abundance, closely related FPs will often appear as well (Bradford et al., 2024; Portik et al., 2022; Sun et al., 2023). These FPs may, by chance, have more misaligned bases than correctly aligned ones for a low-abundance TP, making it difficult to distinguish between the two across all mixes. However, this issue is less likely with template identity, as the value for misaligned reads to a FP tends to be lower than for correctly aligned reads to a low-abundance TP. Moreover, although lower abundances resulted in insufficient reads to fully cover the reference genome of a TP (Fig. 3), the template identities of the TPs were still higher than any FP (Fig. 2). Besides the read length, biases in the database can also influence classification. The most notable biases include the underrepresentation of poorly studied organisms, the overrepresentation of model organisms and clinically relevant species, as well as issued with misannotations and low-quality reference genomes (Chorlton, 2024; Liu et al., 2024). For the Bacilli, we addressed these biases by using a highly curated database for *B. cereus* genomes, BTyperDB, and performing ourselves curation of *B. subtilis* genomes to obtain a high-quality set of reference genomes. When extended to non-*Bacillus* or poorly studies species, this framework may be more strongly influenced by database issues. Additionally, other factors such as similarity of reference genomes to the isolates and the degree of similarity amongst the reference genomes themselves, can impact the classification results (Marcelino et al., 2020; van Bemmelen et al., 2025). Low-abundance species are especially vulnerable to these database effects, as undetected reads or misclassifications have a greater impact when only a few reads are available, in contrast to high-abundance species with much larger read counts. Consequently, using a different the database will likely affect results.

Since the *in silico* mixes did not exactly replicate real sample conditions, both stochastic factors and biases could potentially have been introduced, hence requiring validation with *in vitro* samples essential. Five *in vitro* mixes were designed to mimic real FE contaminations, including low-abundance cases (see Table 3) (D'aes et al., 2025), then length-transformed and downsampled to match *in silico* yields. Analysis with the bioinformatics framework confirmed all TPs had much higher template identities than the FPs, validating the [15.62 %–33.53 %] range to distinguish FPs ($<$15.62 %) from TPs ($\geq$ 33.53 %). Additional analyses on non-transformed *in vitro* samples showed longer reads improved species detection, highlighting the importance of accounting for shorter read length when developing bioinformatics workflows for shotgun metagenomics data.

Finally, the applicability of the framework was assessed using real FE samples. The use of spiked samples was not feasible due to the difficulty of obtaining a clean matrix free of background DNA. Moreover, the spiked microorganism or DNA would not have undergone the same natural degradation processes as real FE samples. Therefore, real FE samples were used instead, which had been exhaustively analyzed using multiple methods in a previous study (D'aes et al., 2025), with the identified Bacilli serving as the ground truth. Applying the developed strategy to the real FE samples confirmed that all present Bacilli species

were accurately detected, provided the combination of sequencing yield (a minimum post-filtering yield of 950 M bases) and relative abundance (at least 1 %) was sufficient.

Traditional methods typically rely on numerous targeted assays for detecting contaminants and conducting follow-up analyses for species detection. Moreover, they require prior information that is typically not available to enforcement laboratories due to the confidential nature of the dossiers submitted to EFSA for authorization of FE on the EU market. The use of shotgun metagenomics bypasses these restrictions to allow detection of *Bacilli* at species level. Distinguishing species within the *Bacillus* genus is challenging due to its complex phylogeny and the high genomic similarity amongst its members. We demonstrated that shotgun metagenomics using nanopore sequencing provides a holistic and culture-free alternative through a single sequencing test. This general open-approach detection method is hence especially suitable for FE products, where identifying Bacilli contaminants, whether production strains or other contaminants, with our validated bioinformatics framework constitutes a first step that can be standardized and harmonized amongst enforcement laboratories. Although our framework is limited to species detection, the availability of shotgun metagenomics data also enables additional types of analyses, such as comprehensive characterization of GM events (White & Hesselberth, 2022), detection of AMR genes (Wang et al., 2024), insights into functional diversity (Waschulin et al., 2022) and more (Zhang et al., 2024). Reads classified as a specific species with KMA can be isolated for further strain-level classification through detailed read mapping and specialized tools such as Floria (Shaw et al., 2024). Additionally, if enough reads are available, metagenomes can be assembled and binned for further strain characterization (D'aes et al., 2025). This can be relevant to investigate the presence of potential AMR and/or toxin genes in wild-type strains, but also follow-up investigation of GM production strains for regulatory purposes. As such, the detection of contaminants with our validated workflow and potential subsequent analyses can be performed all on the same data, reducing the need for other methods. However, for detection of genetically modified species where the GM construct constitutes an episomal plasmid, unambiguous association of the GM construct with the host strain remains difficult without isolation or sample pretreatment such as Hi-C, which crosslinks DNA inside intact cells (Yang et al., 2025). Other studies have shown that such a comprehensive investigation is possible (Buytaers, Fraiture, et al., 2021; D'aes et al., 2025), whereby our validated bioinformatics framework provides a critical first step by allowing accurate species identification to guide potentially warranted follow-up analyses.

We expect the framework to also perform well on other fermentation products, such as heterologous proteins, antibiotics, vitamins and amino acids (Su et al., 2020) due to shared features like Bacilli contaminants, relatively clean sample matrices and similar downstream processing steps. However, for fermentation products produced by other species, such as those from the *Aspergillus* genus (Behera, 2020; Seidler et al., 2024), thresholds defined for Bacilli are unlikely to be directly transferable. Because *Aspergillus* genomes are on average eight times larger than Bacilli (Gibbons & Rokas, 2013) and template identity is across the full genome length, higher sequencing yields will likely be required to clearly separate TPs from FPs. Furthermore, the complexity of fungal genomes complicates unambiguous assignments of reads, but also results in both less reference genomes and lower assembly levels of reference genomes in reference databases. Underrepresentation or incomplete databases increase the chance that reads will map to a wrong, closely related species rather than their correct reference. Lastly, delineating fungal species is particularly complicated (Stengel et al., 2022), making curation efforts more difficult and less straightforward. These considerations indicate that the Bacilli thresholds are not directly applicable and that for fungal contaminants both sensitivity and specificity will be lower. Consequently, establishing appropriate thresholds and LODs for *Aspergillus*, as well as for other species, would require restarting the entire development and evaluation. This process involves defining new *in silico* thresholds using *in silico* mixes of expected contaminants, validating them with *in vitro* mixes at relevant abundances and testing their applicability on real world samples. Additionally, the reference genomes of the specific contaminant(s) in the database should be assessed on taxonomic correctness.

To extend the framework further than fermentation products, one consideration must be acknowledged. FE samples, and by extension fermentation products, governed by strict European legislative standards in the EU, are relatively clean and lack the complex matrices, both chemically and biologically, often found in other sample types. As a result, they undergo post-processing steps that limit residual DNA, allowing for easier species detection when analyzed with shotgun metagenomics. In more complex sample matrices, our developed detection framework may hence not offer the same performance and struggle to distinguish TPs from FPs as effectively due to the increased presence of background noise such as environmental samples (Kuhn et al., 2017) or those with high host fractions (Pereira-Marques et al., 2019). In such matrices, an additional step will likely need to be added, in which host fractions are removed with existing tools to increase the signal-to-noise ratio. Additionally, detection levels will differ for the same yield as the host fraction will take up a part of the sequencing yield, unless adaptive sequencing is used (Martin et al., 2022). Other post-processing steps than used in FE samples will likely also result in different length distributions, which could have an effect on the classification and thus would need to be re-analyzed. Chemical residues, including polysaccharides and secondary metabolites in plants (Schenk et al., 2023) or humic acids and heavy metals in soil, can further complicate detection (Wydro, 2022).

Although the application of shotgun metagenomics in the food industry has expanded over the past decade, the absence of standardized methodologies and the challenges associated with interpreting complex results have hindered its adoption for decision-making in food safety and regulatory practices (Delikanli-Kiyak et al., 2025). While in this study contaminants could be detected up to 1 % with a minimum post-filtering yield of 950 M bases, regulatory frameworks such as employed by EFSA rather define safety criteria in terms of absence of viable cells, expressed in colony-forming units (CFU) per gram (European Food Safety Authority (EFSA), 2021). Since sequence abundance does not provide information on whether the detected species are viable, a direct comparison between the framework's results and safety criteria is hence currently not possible. Consequently, the significance of detecting a contaminant at 1 % relative sequence abundance from a regulatory perspective, along with safety regulations, are not straight-forward and further investigation is needed to clarify how metagenomics can be integrated in regulatory frameworks. It is also important to note that the interpretation of a 1 % relative sequence abundance depends on the total microbial load and would vary if the microbial load input changes. Furthermore, a comparison of sensitivity between shotgun metagenomics and more traditional PCR-based methods is challenging. This is because qPCR results are reported as Ct values, which reflect both absolute copy number and relative abundance (see Supplementary Table S10), whereas shotgun metagenomics directly measures relative abundance through base counts. However, a well-optimized qPCR assay can reliably detect fewer than 25 copies, typically corresponding to Ct values between 36 and 40 (Fraiture et al., 2022, 2023). In combination with using only 10 ng of input compared to 1000 ng for shotgun metagenomics, a higher sensitivity for qPCR is therefore expected. Regarding specificity, qPCR assays must be individually designed and validated, which is a time-consuming process and particularly challenging for *Bacillus* species. Do note that once the qPCR assays are developed, they become significantly more cost-effective than metagenomics. In contrast, shotgun metagenomics, and by extension the framework, offers a distinct advantage by providing higher taxonomic resolution without the need to develop multiple species-specific assays. Consequently, it can be used for risk assessment, serving as an early warning system during the screening of FE products to detect potential

B. cereus contaminations before it enters the food chain. If a positive result is obtained, complementary analyses, such as culture-based enumeration and/or toxin gene detection, should be conducted (Hazards (BIOHAZ), 2016). These complementary analyses help determine whether viable, toxigenic *B cereus* are present. This is especially relevant as the presence of such *B. cereus* in FE can pose a risk due to their potential for proliferation and/or toxin production, for which several examples have been documented, including dairy products (Tirloni et al., 2022), ready-to-eat foods (Yu et al., 2020), fermented soy products (Kim & Mah, 2025) and starch-based foods (Rodrigo et al., 2021). Our framework can also be integrated into quality control procedures for food manufacturers, particularly for FE ingredients. Detection of *B. cereus* enables producers to take preventive measures such as adjusting fermentation parameters or selecting alternative raw materials.

In conclusion, this study shows that shotgun metagenomics in combination with nanopore sequencing can be used to reliably detect Bacilli production strains and contaminants in FE samples, even at low abundances. The framework represents a promising tool for contamination monitoring, supporting robust regulatory standards and quality control practices in FE samples. Furthermore, the automated nature of the framework, combined with the use of standardized interpretation thresholds, enables consistent and user-independent analysis of the output results, although in specific cases a twilight zone can remain that requires manual investigation. Our implemented framework is tailored specifically towards the detection of Bacilli in FE samples, but it can be adapted to other species and fermentation products using a similar *in silico* and *in vitro* approach to evaluate its performance and establish thresholds for reliable species detection.

CRediT authorship contribution statement

Alexander Van Uffelen: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Andrés Posadas:** Writing – review & editing, Data curation. **Marie-Alice Fraiture:** Writing – review & editing, Conceptualization. **Nancy H.C. Roosens:** Writing – review & editing, Project administration, Funding acquisition, Conceptualization. **Sigrid C.J. De Keersmaecker:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Kathleen Marchal:** Writing – review & editing, Methodology, Conceptualization. **Kevin Vanneste:** Writing – review & editing, Supervision, Project administration, Conceptualization.

Code availability

The code used to generate and analyze the data is available at https://github.com/BioinformaticsPlatformWIV-ISP/IsolateMixingAndClassification.

Funding

The research was funded by Sciensano, Belgium (contract METAMORPHOSE).

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fochms.2025.100309.

Data availability

The sequencing data of the isolates and *in vitro* mixes have been deposited in the European Nucleotide Archive under accession number PRJEB87256. Data for the *in silico* mixes is not available, but can be generate through the scripts under Code Availability. The data of the commercial FE samples is available under the accessions in Supplementary Table S1. The NCBI and BTyperDB accessions of the sequences in the database are available on Zenodo (doi: https://doi.org/10.5281/zenodo.15388579).

References

Behera, B. C. (2020). Citric acid from *Aspergillus niger*: A comprehensive overview. *Critical Reviews in Microbiology, 46*(6), 727–749. https://doi.org/10.1080/1040841X.2020.1828815

van Bemmelen, J., Nika, I., & Baaijens, J. A. (2025). Benchmarking the impact of reference genome selection on taxonomic profiling accuracy [Preprint]. *Bioinformatics*. https://doi.org/10.1101/2025.02.07.637076

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. https://doi.org/10.3389/fmicb.2023.1173594

Bradford, L. M., Carrillo, C., & Wong, A. (2024). Managing false positives during detection of pathogen sequences in shotgun metagenomics datasets. *BMC Bioinformatics, 25*(1), 372. https://doi.org/10.1186/s12859-024-05952-x

Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2019). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics, 20*(4), 1125–1136. https://doi.org/10.1093/bib/bbx120

Buytaers, F. E., Fraiture, M.-A., Berbers, B., Vandermassen, E., Hoffman, S., Papazova, N., … De Keersmaecker, S. C. J. (2021). A shotgun metagenomics approach to detect and characterize unauthorized genetically modified microorganisms in microbial fermentation products. *Food Chemistry: Molecular Sciences, 2*, Article 100023. https://doi.org/10.1016/j.fochms.2021.100023

Buytaers, F. E., Saltykova, A., Denayer, S., Verhaegen, B., Vanneste, K., Roosens, N. H. C., … De Keersmaecker, S. C. J. (2021). Towards real-time and affordable strain-level metagenomics-based foodborne outbreak investigations using Oxford Nanopore Sequencing Technologies. *Frontiers in Microbiology, 12*, Article 738284. https://doi.org/10.3389/fmicb.2021.738284

Carroll, L. M., Cheng, R. A., Wiedmann, M., & Kovac, J. (2022). Keeping up with the *Bacillus cereus* group: Taxonomy through the genomics era and beyond. *Critical Reviews in Food Science and Nutrition, 62*(28), 7677–7702. https://doi.org/10.1080/10408398.2021.1916735

Chen, Y., Li, Y., Shen, J., Liu, Q., Liu, Y., Chu, Y., & Xiao, Z. (2022). *Bacillus arachidis* sp. nov., isolated from peanut rhizosphere soil. *Current Microbiology, 79*(8), 231. https://doi.org/10.1007/s00284-022-02925-2

Chorlton, S. D. (2024). Ten common issues with reference sequence databases and how to mitigate them. *Frontiers in Bioinformatics, 4*, Article 1278228. https://doi.org/10.3389/fbinf.2024.1278228

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*(1), 307. https://doi.org/10.1186/s12859-018-2336-6

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*(12), 1971. https://doi.org/10.3390/life12121971

D'aes, J., Fraiture, M.-A., Bogaerts, B., Laere, Y. V., Keersmaecker, S. C. J. D., Roosens, N. H. C., & Vanneste, K. (2025). Metagenomics-based tracing of genetically modified microorganism contaminations in commercial fermentation products. *Food Chemistry: Molecular Sciences, 10*, Article 100236. https://doi.org/10.1016/j.fochms.2024.100236

Danilova, I., & Sharipova, M. (2020). The practical potential of Bacilli and their enzymes for industrial production. *Frontiers in Microbiology, 11*, 1782. https://doi.org/10.3389/fmicb.2020.01782

Davis, B. C., Brown, C., Gupta, S., Calarco, J., Liguori, K., Milligan, E., … Keenum, I. (2023). Recommendations for the use of metagenomics for routine monitoring of antibiotic resistance in wastewater and impacted aquatic environments. *Critical Reviews in Environmental Science and Technology, 53*(19), 1731–1756. https://doi.org/10.1080/10643389.2023.2181620

Deckers, M., Vanneste, K., Winand, R., Keersmaecker, S. C. J. D., Denayer, S., Heyndrickx, M., … Roosens, N. H. C. (2020). Strategy for the identification of micro-organisms producing food and feed products: Bacteria producing food enzymes as study case. *Food Chemistry, 305*, Article 125431. https://doi.org/10.1016/j.foodchem.2019.125431

Delikanli-Kiyak, B., Yilmaz, I., & Guldas, M. (2025). Can metagenomic analyses be used effectively in safe food production? *Food Science & Nutrition, 13*(8), Article e70772. https://doi.org/10.1002/fsn3.70772

Du, Y., Ma, J., Yin, Z., Liu, K., Yao, G., Xu, W., … Wang, C. (2019). Comparative genomic analysis of Bacillus paralicheniformis MDJK30 with its closely related species reveals an evolutionary relationship between *B. paralicheniformis* and *B. licheniformis*. *BMC Genomics, 20*(1), 283. https://doi.org/10.1186/s12864-019-5646-9

Dunlap, C. A., Bowman, M. J., & Zeigler, D. R. (2019). Promotion of Bacillus subtilis subsp. inaquosorum, *Bacillus subtilis* subsp. spizizenii and Bacillus subtilis subsp. stercoris to species status. *Antonie Van Leeuwenhoek, 113*(1), 1–12. https://doi.org/10.1007/s10482-019-01354-9

EFSA Panel on Food Contact Materials, Enzymes and Processing Aids (CEP), Barat Baviera, J. M., Bolognesi, C., Chesson, A., Cocconcelli, P. S., … Peluso, S. (2024). Taxonomic identity of the Bacillus licheniformis strains used to produce food enzymes evaluated in published EFSA opinions. *EFSA Journal, 22*(5). https://doi.org/10.2903/j.efsa.2024.8770

Ehling-Schulz, M., Lereclus, D., & Koehler, T. M. (2019). The *Bacillus cereus* group: *Bacillus* species with pathogenic potential. *Microbiology Spectrum, 7*(3), 7.3.6. https://doi.org/10.1128/microbiolspec.GPP3-0032-2018

European Food Safety Authority (EFSA). (2021). Scientific Guidance for the submission of dossiers on Food Enzymes. *EFSA Journal, 19*(10). https://doi.org/10.2903/j.efsa.2021.6851

European Parliament & Council of the European Union. (2008). Regulation (EC) No 1332/2008 of the European Parliament and of the Council. *Official Journal of the European Union, 354*, 7–15.

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*(1), 7094. https://doi.org/10.1038/s41598-020-63987-5

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. https://doi.org/10.1016/j.ijfoodmicro.2020.108749

Fraiture, M.-A., Gobbo, A., Papazova, N., & Roosens, N. H. C. (2022). Development of a taxon-specific real-time PCR method targeting the *Bacillus subtilis* group to strengthen the control of genetically modified bacteria in fermentation products. *Fermentation, 8*(2), 78. https://doi.org/10.3390/fermentation8020078

Fraiture, M.-A., Gobbo, A., Papazova, N., & Roosens, N. H. C. (2023). Development of a taxon-specific real-time polymerase chain reaction method to detect *Trichoderma reesei* contaminations in fermentation products. *Fermentation, 9*(11), 926. https://doi.org/10.3390/fermentation9110926

Gand, M., Navickaite, I., Bartsch, L.-J., Grützke, J., Overballe-Petersen, S., Rasmussen, A., … De Keersmaecker, S. C. J. (2024). Towards facilitated interpretation of shotgun metagenomics long-read sequencing data analyzed with KMA for the detection of bacterial pathogens and their antimicrobial resistance genes. *Frontiers in Microbiology, 15*, Article 1336532. https://doi.org/10.3389/fmicb.2024.1336532

Gibbons, J. G., & Rokas, A. (2013). The function and evolution of the aspergillus genome. *Trends in Microbiology, 21*(1), 14–22.

Govender, K. N., & Eyre, D. W. (2022). Benchmarking taxonomic classifiers with Illumina and Nanopore sequence data for clinical metagenomic diagnostic applications. *Microbial Genomics, 8*(10). https://doi.org/10.1099/mgen.0.000886

Hazards (BIOHAZ), E. P. on B. (2016). Risks for public health related to the presence of *Bacillus cereus* and other Bacillus spp. including *Bacillus thuringiensis* in foodstuffs. *EFSA Journal, 14*(7). https://doi.org/10.2903/j.efsa.2016.4524

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*(1), 5114. https://doi.org/10.1038/s41467-018-07641-9

Jovanovic, J., Ornelis, V. F. M., Madder, A., & Rajkovic, A. (2021). *Bacillus cereus* food intoxication and toxicoinfection. *Comprehensive Reviews in Food Science and Food Safety, 20*(4), 3719–3761. https://doi.org/10.1111/1541-4337.12785

Kenfaoui, J., Dutilloy, E., Benchlih, S., Lahlali, R., Ait-Barka, E., & Esmaeel, Q. (2024). *Bacillus velezensis*: A versatile ally in the battle against phytopathogens—Insights and prospects. *Applied Microbiology and Biotechnology, 108*(1), 439. https://doi.org/10.1007/s00253-024-13255-7

Kim, C., Pongpanich, M., & Porntaveetus, T. (2024). Unraveling metagenomics through long-read sequencing: A comprehensive review. *Journal of Translational Medicine, 22*(1), 111. https://doi.org/10.1186/s12967-024-04917-1

Kim, K. H., Han, D. M., Lee, J. K., & Jeon, C. O. (2023). Alkalicoccobacillus porphyridii sp. nov., isolated from a marine red alga, reclassification of Shouchella plakortidis and Shouchella gibsonii as Alkalicoccobacillus plakortidis comb. nov. and Alkalicoccobacillus gibsonii comb. nov., and emended description of the genus Alkalicoccobacillus Joshi et al. 2022. *International Journal of Systematic and Evolutionary Microbiology, 73*(8). https://doi.org/10.1099/ijsem.0.006019

Kim, S., & Mah, J.-H. (2025). Variation in heat resistance and biofilm formation of *Bacillus cereus* spores in various fermented soybean foods. *International Journal of Food Microbiology, 427*, Article 110939. https://doi.org/10.1016/j.ijfoodmicro.2024.110939

Kuhn, R., Böllmann, J., Krahl, K., Bryant, I. M., & Martienssen, M. (2017). Comparison of ten different DNA extraction procedures with respect to their suitability for environmental samples. *Journal of Microbiological Methods, 143*, 78–86. https://doi.org/10.1016/j.mimet.2017.10.007

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics, 34*(18), 3094–3100.

Liu, Y., Ghaffari, M. H., Ma, T., & Tu, Y. (2024). Impact of database choice and confidence score on the performance of taxonomic classification using Kraken2. *ABIOTECH, 5*(4), 465–475. https://doi.org/10.1007/s42994-024-00178-0

Marcelino, R., Holmes, E. C., & Sorrell, T. C. (2020). The use of taxon-specific reference databases compromises metagenomic classification. *BMC Genomics, 21*(1), 184. https://doi.org/10.1186/s12864-020-6592-2

Martin, S., Heavens, D., Lan, Y., Horsfield, S., Clark, M. D., & Leggett, R. M. (2022). Nanopore adaptive sampling: A tool for enrichment of low abundance species in metagenomic samples. *Genome Biology, 23*(1). https://doi.org/10.1186/s13059-021-02582-x

O'Leary, N. A., Cox, E., Holmes, J. B., Anderson, W. R., Falk, R., Hem, V., … Schneider, V. A. (2024). Exploring and retrieving sequence and metadata for species across the tree of life with NCBI datasets. *Scientific Data, 11*(1), 732. https://doi.org/10.1038/s41597-024-03571-y

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research, 44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology, 17*(1), 132. https://doi.org/10.1186/s13059-016-0997-x

Oxford Nanopore PLC. (2023). Dorado. In *[GitHub repository] (https://github.com/nanoporetech/dorado)*. GitHub. https://github.com/nanoporetech/dorado.

Paracchini, V., Petrillo, M., Reiting, R., Angers-Loustau, A., Wahler, D., Stolz, A., … Grohmann, L. (2017). Molecular characterization of an unauthorized genetically modified *Bacillus subtilis* production strain identified in a vitamin B 2 feed additive. *Food Chemistry, 230*, 681–689. https://doi.org/10.1016/j.foodchem.2017.03.042

Pereira-Marques, J., Hout, A., Ferreira, R. M., Weber, M., Pinto-Ribeiro, I., van Doorn, L.-J., … Figueiredo, C. (2019). Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis. *Frontiers in Microbiology, 10*, 1277. https://doi.org/10.3389/fmicb.2019.01277

Portik, D. M., Brown, C. T., & Pierce-Ward, N. T. (2022). Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets. *BMC Bioinformatics, 23*(1), 541. https://doi.org/10.1186/s12859-022-05103-0

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology, 35*(9), 833–844. https://doi.org/10.1038/nbt.3935

Rahnama, H., Azari, R., Yousefi, M. H., Berizi, E., Mazloomi, S. M., Hosseinzadeh, S., … Conti, G. O. (2023). A systematic review and meta-analysis of the prevalence of *Bacillus cereus* in foods. *Food Control, 143*, Article 109250. https://doi.org/10.1016/j.foodcont.2022.109250

Ramnath, V., Larralde, M., Menchik, P., Buehler, A. J., Harrand, A. S., Chung, T., … Carroll, L. M. (2023). A community-curated, global atlas of *Bacillus cereus* sensu lato genomes for epidemiological surveillance [Preprint]. *Microbiology*. https://doi.org/10.1101/2023.12.20.572685

Rodrigo, D., Rosell, C. M., & Martinez, A. (2021). Risk of *Bacillus cereus* in relation to rice and derivatives. *Foods, 10*(2), 302. https://doi.org/10.3390/foods10020302

Schenk, J. J., Becklund, L. E., Carey, S. J., & Fabre, P. P. (2023). What is the "modified" CTAB protocol? Characterizing modifications to the CTAB DNA extraction protocol. *Applications in Plant Sciences, 11*(3), Article e11517. https://doi.org/10.1002/aps3.11517

Seidler, Y., Rimbach, G., Lüersen, K., Vinderola, G., & Ipharraguerre, I. R. (2024). The postbiotic potential of *Aspergillus oryzae* – A narrative review. *Frontiers in Microbiology, 15*, Article 1452725. https://doi.org/10.3389/fmicb.2024.1452725

Shaw, J., Gounot, J.-S., Chen, H., Nagarajan, N., & Yu, Y. W. (2024). Floria: Fast and accurate strain haplotyping in metagenomes. *Bioinformatics, 40*(Supplement_1), i30–i38. https://doi.org/10.1093/bioinformatics/btae252

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One, 11*(10), Article e0163962. https://doi.org/10.1371/journal.pone.0163962

Simon, H. Y., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking metagenomics tools for taxonomic classification. *Cell, 178*(4), 779–794.

Stenfors Arnesen, L. P., Fagerlund, A., & Granum, P. E. (2008). From soil to gut: *Bacillus cereus* and its food poisoning toxins. *FEMS Microbiology Reviews, 32*(4), 579–606. https://doi.org/10.1111/j.1574-6976.2008.00112.x

Stengel, A., Stanke, K. M., Quattrone, A. C., & Herr, J. R. (2022). Improving taxonomic delimitation of fungal species in the age of genomics and Phenomics. *Frontiers in Microbiology, 13*. https://doi.org/10.3389/fmicb.2022.847067

Strube, M. L. (2021). RibDif: Can individual species be differentiated by 16S sequencing? *Bioinformatics Advances, 1*(1), Article vbab020. https://doi.org/10.1093/bioadv/vbab020

Su, T., Shen, B., Hu, X., Teng, Y., Weng, P., Wu, Z., & Liu, L. (2024). Research advance of *Bacillus velezensis*: Bioinformatics, characteristics, and applications. *Food Science and Human Wellness, 13*(4), 1756–1766. https://doi.org/10.26599/FSHW.2022.9250148

Su, Y., Liu, C., Fang, H., & Zhang, D. (2020). *Bacillus subtilis*: A universal cell factory for industry, agriculture, biomaterials and medicine. *Microbial Cell Factories, 19*(1), 173. https://doi.org/10.1186/s12934-020-01436-8

Sun, Z., Liu, J., Zhang, M., Wang, T., Huang, S., Weiss, S. T., & Liu, Y.-Y. (2023). Removal of false positives in metagenomics-based taxonomy profiling via targeting type IIB restriction sites. *Nature Communications, 14*(1), 5321. https://doi.org/10.1038/s41467-023-41099-8

Taxt, A. M., Avershina, E., Frye, S. A., Naseer, U., & Ahmad, R. (2020). Rapid identification of pathogens, antibiotic resistance genes and plasmids in blood cultures by nanopore sequencing. *Scientific Reports, 10*(1), 7622. https://doi.org/10.1038/s41598-020-64616-x

Tirloni, E., Stella, S., Celandroni, F., Mazzantini, D., Bernardi, C., & Ghelardi, E. (2022). *Bacillus cereus* in dairy products and production plants. *Foods, 11*(17), 2572. https://doi.org/10.3390/foods11172572

Van Uffelen, A., Posadas, A., Roosens, N. H. C., Marchal, K., De Keersmaecker, S. C. J., & Vanneste, K. (2024). Benchmarking bacterial taxonomic classification using

nanopore metagenomics data of several mock communities. *Scientific Data, 11*(1), 864. https://doi.org/10.1038/s41597-024-03672-8

Wang, Y., Xu, N., Chen, B., Zhang, Z., Lei, C., Zhang, Q., … Qian, H. (2024). Metagenomic analysis of antibiotic-resistance genes and viruses released from glaciers into downstream habitats. *Science of the Total Environment, 908*, Article 168310. https://doi.org/10.1016/j.scitotenv.2023.168310

Waschulin, V., Borsetto, C., James, R., Newsham, K. K., Donadio, S., Corre, C., & Wellington, E. (2022). Biosynthetic potential of uncultured Antarctic soil bacteria revealed through long-read metagenomic sequencing. *The ISME Journal, 16*(1), 101–111. https://doi.org/10.1038/s41396-021-01052-3

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software, 6*(60), 3021. https://doi.org/10.21105/joss.03021

White, L. K., & Hesselberth, J. R. (2022). Modification mapping by nanopore sequencing. *Frontiers in Genetics, 13*, Article 1037134. https://doi.org/10.3389/fgene.2022.1037134

Wydro, U. (2022). Soil microbiome study based on DNA extraction: A review. *Water, 14*(24), 3999. https://doi.org/10.3390/w14243999

Xu, X., & Kovács, Á. T. (2024). How to identify and quantify the members of the *Bacillus* genus? *Environmental Microbiology, 26*(2), Article e16593. https://doi.org/10.1111/1462-2920.16593

Yahara, K., Suzuki, M., Hirabayashi, A., Suda, W., Hattori, M., Suzuki, Y., & Okazaki, Y. (2021). Long-read metagenomics using PromethION uncovers oral bacteriophages and their interaction with host bacteria. *Nature Communications, 12*(1), 27. https://doi.org/10.1038/s41467-020-20199-9

Yang, Q. E., Gao, J. T., Zhou, S. G., & Walsh, T. R. (2025). Cutting-edge tools for unveiling the dynamics of plasmid–host interactions. *Trends in Microbiology., 33*(5), 496–509. https://doi.org/10.1016/j.tim.2024.12.013

Yu, S., Yu, P., Wang, J., Li, C., Guo, H., Liu, C., … Ding, Y. (2020). A study on prevalence and characterization of *Bacillus cereus* in ready-to-eat foods in China. *Frontiers in Microbiology, 10*. https://doi.org/10.3389/fmicb.2019.03043

Zhang, T., Li, H., Jiang, M., Hou, H., Gao, Y., Li, Y., … Liu, Y.-X. (2024). Nanopore sequencing: Flourishing in its teenage years. *Journal of Genetics and Genomics, 51*(12), 1361–1374. https://doi.org/10.1016/j.jgg.2024.09.007

Zhang, Y., Geary, T., & Simpson, B. K. (2019). Genetically modified food enzymes: A review. *Current Opinion in Food Science, 25*, 14–18. https://doi.org/10.1016/j.cofs.2019.01.002