Food Chemistry: Molecular Sciences 10 (2025) 100236

Contents lists available at ScienceDirect

Food Chemistry: Molecular Sciences

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

Metagenomics-based tracing of genetically modified microorganism contaminations in commercial fermentation products

Jolien D'aes[a,1], Marie-Alice Fraiture[a,1], Bert Bogaerts[a], Yari Van Laere[a,b], Sigrid C.J. De Keersmaecker[a], Nancy H.C. Roosens[a,2], Kevin Vanneste[a,*,2]

[a] Sciensano, Transversal activities in Applied Genomics (TAG), J. Wytsmanstraat 14, 1050 Brussels, Belgium
[b] UGent, Department of Plant Biotechnology & Bioinformatics, Technologiepark 71 9052 Zwijnaarde, Belgium

ARTICLE INFO

Keywords:
Food enzyme
Phylogenomics
Strain-level deconvolution
Antimicrobial resistance

ABSTRACT

Genetically modified microorganisms (GMM) are frequently employed for the production of microbial fermentation products such as food enzymes. Although presence of the GMM or its recombinant DNA in the final product is not authorized, contaminations occur frequently. Insight into the contamination source of a GMM is of crucial importance to allow the competent authorities to take appropriate action. The aim of this study was to explore the feasibility of a metagenomic shotgun sequencing approach to investigate microbial contamination in fermentation products, focusing on source tracing of GMM strains using innovative strain deconvolution and phylogenomic approaches. In most cases, analysis of 16 GMM-contaminated food enzyme products supported finding the same GM producer strains in different products, while often multiple GMM contaminations per product were detected. Presence of AMR genes in the samples was strongly associated with GMM contamination, emphasizing the potential public health risk. Additionally, a variety of other microbial contaminations were detected, identifying a group of samples with a conspicuously similar contamination profile, which suggested that these samples originated from the same production facility or batch. Together, these findings highlight the need for guidelines and quality control for traceability of these products to ensure the safety of consumers. This study demonstrates the added value of metagenomics to obtain insight in the microbial contamination profiles, as well as their underlying relationships, in commercial microbial fermentation products. The proposed approach may be applied to other types of microbial fermentation products and/or to other (genetically modified) producer strains.

1. Introduction

Naturally fermented foods, such as cheese, yoghurt and tofu, are a valuable component of the human diet. Their production depends on the activity of micro-organisms that digest natural substrates, converting them into enzymes or metabolites that play a key role in the taste and properties of the final product. On the other hand, many food supplements such as vitamins, additives, or enzymes, with important applications in the food and feed industry, are the result of industrial fermentation with axenic microbial cultures (Graham & Ledesma-Amaro, 2023). These microbial fermentation products are increasingly being produced by genetically modified microorganisms (GMM), mainly because of the increases in productivity and yield during the fermentation process (Deckers, Deforce, et al., 2020).

Although novel gene editing techniques, such as CRISPR/Cas technology, allow for swift and targeted introduction of unmarked genetic modifications, in practice many GMM constructs still incorporate antimicrobial resistance (AMR) genes. The selection pressure that can be imposed by the presence of these genes is often needed to prevent loss of the high-copy plasmid or multi-copy chromosomal insertion carrying the transgenic construct. However, a public health risk may arise if a GMM carrying AMR genes contaminates the final fermented product, since those AMR genes could spread via horizontal gene transfer to other bacteria upon ingestion, particularly if residing on a plasmid (Arnold et al., 2022; Florez-Cuadrado et al., 2018; Von Wintersdorff et al., 2016).

The presence of a genetically modified organism (GMO) in a product

[*] Corresponding author.
E-mail address: kevin.vanneste@sciensano.be (K. Vanneste).
[1] equal first-author contribution
[2] equal last-author contribution

https://doi.org/10.1016/j.fochms.2024.100236
Received 6 September 2024; Received in revised form 16 December 2024; Accepted 19 December 2024
Available online 24 December 2024

destined for food or feed purposes, whether as a living strain or associated recombinant DNA, requires prior authorization in line with EU regulations (EC 1830/2003). Currently, no authorizations have been granted for GMM intended for the food or feed chain on the EU market. Thus, a GMM that was used to generate a fermentation product must be absent from the final product destined for consumption. To ensure traceability and food safety, a qPCR-based GMM detection strategy was developed for enforcement laboratories to support the competent authorities. Within this scope, qPCR assays targeting the unnatural associations of several GMM constructs were designed, enabling their identification in the final product (Fraiture, Deckers, et al., 2020a; Fraiture, Deckers, et al., 2020b; Fraiture, Deckers, et al., 2020c; Fraiture, Gobbo, et al., 2021; Fraiture, Marchesi, et al., 2021; Fraiture, Papazova, & Roosens, 2021). As these identification assays focus on the transgenic construct rather than the host strain of the GMM, it cannot be ruled out that different host strains, carrying a similar (episomal) plasmid construct, may be present in different samples. Moreover, GMM constructs for which no qPCR assay is available, will be missed entirely.

Using these qPCR-based screening methods, previous research has demonstrated GMM contaminations in a variety of food enzyme (FE) products, leading to 18 RASFF notifications since 2019 (https://ec.europa.eu/food/safety/rasff-food-and-feed-safety-alerts/rasff-portal_en). Additionally, a viable genetically modified (GM) *Bacillus velezensis* strain was isolated from several FE products, and characterized by whole-genome sequencing (WGS) (D'aes et al., 2021; Fraiture, Bogaerts, et al., 2020). This strain, designated GMM 'protease1', harbored the high-copy plasmid vector pUB110 carrying two AMR genes encoding kanamycin/neomycin (*aadD1)* and bleomycin (*bleO*) resistance, alongside an insertion of a protease-encoding gene. The GMM host strains found in the various FE samples were phylogenetically closely linked and probably descended from the same strain, indicating a shared origin of the contaminations in different commercial samples (D'aes et al., 2021). In a subsequent study (D'aes et al., 2022), four alpha-amylase and protease FE products were subjected to metagenomic high-throughput sequencing (HTS), resulting in the characterization of two additional GM strains, which could not be isolated from any of the products. Firstly, a GMM 'amylase1' was characterized that carried the same high-copy pUB110 vector with an insertion of an alpha-amylase encoding gene (*amyA*) derived from *B. amyloliquefaciens*. As the transgenic construct was located on a plasmid, the metagenomic approach could not unequivocally determine its host, although results indicated that it was most likely a *B. amyloliquefaciens* strain. Secondly, a novel GMM 'amylase2' was detected and characterized, constituting a *B. licheniformis* strain carrying a chromosomally integrated transgenic construct comprising an AMR gene encoding resistance to chloramphenicol (*catA*) and an alpha-amylase encoding gene (*amyS*) derived from its *B. licheniformis* host. Both the *B. amyloliquefaciens* and *B. licheniformis* GM strains exhibited signs of genetic modifications in sporulation genes, consisting of deletions that presumably rendered the strains asporogenic. This would make them incapable of long-term survival in the product, which could explain why it was impossible to obtain viable isolates.

Preliminary findings, based on whole-genome comparison of the metagenome-assembled genomes (MAGs) retrieved in the previous study (D'aes et al., 2022), suggested the possibility of a shared origin of the unculturable GMM contaminations. A phylogenetic comparison of the strains found in different FE samples was however not explored, since viable isolates could only be obtained for one of the three detected GMM strains, and performing source tracing based on metagenomics data for the two other GMM strains, constitutes a much more complex challenge. For instance, *B. velezensis* and *B. amyloliquefaciens* are closely related species and cannot always be distinguished in metagenomic samples using 'standard' metagenomic methods, e.g. metagenomic de novo assembly (Fan et al., 2017).

The goal of the present study was to determine the feasibility of a metagenomic approach to gain insight into the contamination origin of microbial fermentation products, which is of crucial importance to allow the competent authorities to take appropriate actions. More specifically, this study aimed to compare microbial contaminants in microbial fermentation products at the strain level, with special emphasis on GMM, to identify any common origin or relationship among them through SNP-based phylogenomic comparisons.

2. Material and methods

2.1. DNA extraction from FE matrix

From each FE product, collected on the EU market (Table 1), a sample of 200 mg was used for genomic DNA extraction using the Quick-DNA™ HMW MagBead Kit (ZymoResearch, Freiburg, Germany) according to the manufacturer's instructions. Extracted DNA was visualized by capillary electrophoresis using the Tapestation 4200 device with the associated genomic DNA Screen Tape and reagents (Agilent, Santa Clara, USA). Each DNA concentration was measured by spectrophotometry using the Nanodrop® 2000 (ThermoFisher, Dilbeek, Belgium), and each DNA purity was evaluated using the A260/A280 and A260/A230 ratios.

2.2. qPCR assays

DNA from FE products was analyzed using qPCR methods specific to the *Bacillus subtilis* group (BSG) (Fraiture et al., 2022), a GM *B. velezensis* producing protease (GMM protease1) (Fraiture, Bogaerts, et al., 2020), a second GMM with a transgenic construct encoding a protease (GMM protease2) (Fraiture, Gobbo, et al., 2021), a GM *B. amyloliquefaciens* producing alpha-amylase (GMM alpha-amylase1) (Fraiture, Marchesi, et al., 2021), and a GM *B. licheniformis* producing alpha-amylase (GMM alpha-amylase2) (Fraiture et al., 2024). qPCR assays were performed in duplicate as described before (D'aes et al., 2022).

2.3. DNA library preparation and sequencing

For some samples (Table S1), metagenomic Illumina sequencing data was already available (D'aes et al., 2022). For the remaining samples, one short-read DNA library per sample was prepared using the Nextera XT DNA library preparation kit (Illumina, California, USA) according to the manufacturer's instructions. Sequencing was carried out on an Illumina MiSeq system with the V3 chemistry, obtaining 250 bp paired-end reads. For most samples, approximately seven FE sample libraries were analyzed together on a MiSeq, in equimolar quantities. Additionally, an entire independent MiSeq run had been previously devoted to sequencing the A3 sample library to obtain a super-high depth sequencing coverage (D'aes et al., 2022).

For some samples (Table S1), ONT sequencing data was already available (D'aes et al., 2022), which was included in this study, while for sample A12 ONT sequencing was performed in this study. One DNA library per sample was prepared using the ligation sequencing kit (SQK-LSK109; Oxford Nanopore Technologies, Oxford, UK) according to the manufacturer's instructions. The FE sample library was loaded on an individual R9 MinION flow cell and sequenced for 48 h. The ONT reads were basecalled with Guppy 5.0.7 in GPU mode, with the dna_r9.4.1_450bps_sup model, and with q-score based filtering disabled.

2.4. Metagenomic approach to characterize and compare microbial contaminations

To characterize and compare the microbial contamination in the samples, different complementary approaches were used (Fig. 1), a detailed description of which is provided in Supplementary text S1.

2.4.1. Raw read preprocessing and analysis

The overall microbial composition of the short-read samples was

Table 1
Overview of samples included in this study.

Sample	Brand	Labeled enzymes	Labeled producer organism	GMM alpha-amylase1	GMM alpha-amylase2	GMM protease1	RASFF
A1	A	alpha-amylase	Bacteria	++	++	++	RASFF2020.2846
A2	B	alpha-amylase	Unknown	++	++	++	RASFF2020.2577
A3	C	alpha-amylase	Unknown	++	++	++	RASFF2020.2577
A4	D	alpha-amylase	Bacteria	++	++	-	RASFF2020.2579
A5	E	alpha-amylase	*Bacillus licheniformis*	++	++	+	RASFF2019.3332
P1	E	protease	*Bacillus subtilis*	++	++	++	RASFF2019.3332
A6	F	alpha-amylase	*Bacillus licheniformis*	-	+	-	RASFF2020.2576
A7	G	alpha-amylase	*Bacillus licheniformis*	+	+	-	/
M1	E	protease, cellulase, xylanase, alpha-amylase, beta-glucanase	*Aspergillus oryzae, Bacillus subtilis, Trichoderma reesei, Trichoderma longibrachiatum*	++	++	++	RASFF2019.3332
P2	H	protease	*Bacillus licheniformis*	+	-	+	/
A8	I	alpha-amylase	Unknown	+	+	-	/
A9	G	alpha-amylase	*Bacillus subtilis*	++	+	-	/
A10	G	alpha-amylase	*Bacillus amyloliquefaciens*	++	-	-	/
A11	J	alpha-amylase	Unknown	++	-	-	RASFF2020.2570
A12	K	alpha-amylase	*Bacillus subtilis*	++	-	-	/
A13	unknown	alpha-amylase	Unknown	++	-	-	RASFF2019.3332

qPCR results with a Cq below 25 were interpreted as 'highly contaminated' (++), below 35 as 'contaminated' (+), and above 35 as 'negative' (-). GMM protease2 was not detected in any of the samples. Exact Cq values are provided in Table S1. Sample names and brands were replaced with aliases for confidentiality reasons. Sample aliases starting with A, P, or M refer to alpha-amylase, protease or mixed enzyme products, respectively. The Rapid Alert System for Food and Feed (RASFF) entries indicate EU-level notifications of potential health risks derived from food or feed.

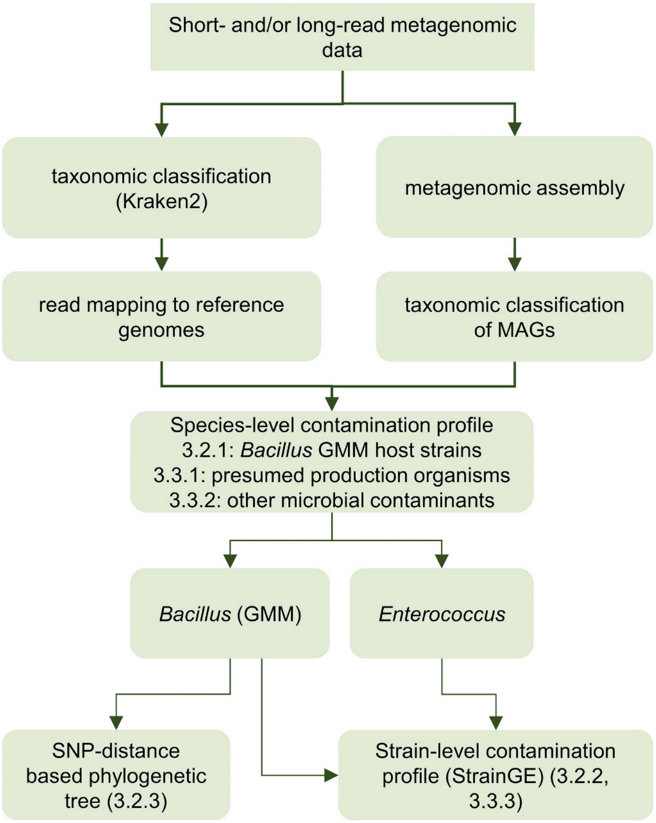

Short- and/or long-read metagenomic data
taxonomic classification (Kraken2)
metagenomic assembly
read mapping to reference genomes
taxonomic classification of MAGs
Species-level contamination profile
3.2.1: *Bacillus* GMM host strains
3.3.1: presumed production organisms
3.3.2: other microbial contaminants
Bacillus (GMM)
Enterococcus
SNP-distance based phylogenetic tree (3.2.3)
Strain-level contamination profile (StrainGE) (3.2.2, 3.3.3)

Fig. 1. Overview of workflow to characterize and compare microbial contaminations in food enzyme samples, with reference to the related results sections.

profiled with Kraken2 (Wood et al., 2019), with a database composed of RefSeq assemblies (retrieved on 24/02/2023) from archaea, bacteria, fungi, protozoa and viruses, complemented with the human genome (Fig. 1). The Kraken2 relative abundance estimates were corrected with Bracken 2.8 (Lu et al., 2017), and visualized with the Python package Seaborn (Waskom, 2021).

Since k-mer-based classifiers such as Kraken2 can potentially be subject to false positive identifications, especially at low abundance levels (Meyer et al., 2022), a complementary read-mapping analysis was performed for every species reported at a relative abundance of at least 1 %. Following read preprocessing with Trimmomatic 0.38 (Bolger et al., 2014), short and long reads were mapped with Bowtie2 2.4.1 (Langmead & Salzberg, 2012), or minimap2 2.26 (Li, 2018), respectively, after which the alignments were analyzed with samtools 1.17 (Li et al., 2009) to calculate the depth and breadth of coverage.

The presence of a species in a sample was considered as confirmed if it was detected by short- and/or long-read mapping to the reference genome of the species with a depth of coverage of at least 5 and a breadth of coverage of at least 80 %, and/or if it was represented by a de novo metagenome assembled genome (MAG) that was at least 50 % complete and contained maximum 10 % contamination (2.4.2, Table S3), i.e. the MAG quality cutoff for reliable taxonomic classification suggested by Chaumeil et al. (2020).

2.4.2. Metagenomic assembly and characterization

In case both Illumina and ONT reads were available for a sample, metagenomic hybrid assembly was carried out with OPERA-MS 0.9.0 (Bertrand et al., 2019), with SPAdes 3.13.0 (Bankevich et al., 2012), Samtools 0.1.19, BWA 0.7.10-r789 (Li & Durbin, 2009), Blasr 5.1, minimap2 2.11-r797, Racon 0.5.0 (Vaser et al., 2017), Mash 2.2 (Ondov et al., 2016), MUMmer 3.23 (Kurtz et al., 2004), and Pilon 1.22 (Walker et al., 2014) as dependencies. If only Illumina reads were available, short-read metagenomic assembly was performed with SPAdes 3.15.3 (Nurk et al., 2017), followed by binning with MetaBAT2 2.15 (Kang et al., 2019).

Completeness and contamination rates of bacterial MAGs were estimated with CheckM 1.1.3 (Parks et al., 2015), with Prodigal 2.6.3 (Hyatt et al., 2010) and pplacer 1.1.alpha19 (Matsen et al., 2010) as dependencies. For bacterial taxonomic classification, GTDB-Tk 1.5.1 (Chaumeil et al., 2020) was employed, with FastANI 1.33 (Jain et al., 2018), FastTree 2.1.11 (Price et al., 2009), Mash 2.2, Prodigal 2.6.3, pplacer 1.1.alpha19, and HMMER 3.2.1 as dependencies. Quality control of fungal MAGs was done with BUSCO 5.4.3 (Manni et al., 2021), with as dependencies Augustus 3.3.3 (Stanke et al., 2008), Metaeuk 6 (Levy Karin et al., 2020), SEPP 4.5.2 (Warnow, 2013), BBtools 38.34 (Bushnell et al., 2017), Prodigal 2.6.3, HMMER 3.2.1, and BLAST+ 2.13.0 (Camacho et al., 2009). To obtain a taxonomic classification of the fungal MAGs, the top hits of a web-based blastn search with default parameters were examined manually.

2.4.3. StrainGE analysis

To investigate a potential shared origin of the GMM contaminations, a strain-level analysis with StrainGE (van Dijk et al., 2022), using the short read datasets, was performed for the *Bacillus* species. Additionally, StrainGE was run for the *Enterococcus* genus, because this was the main microbial contamination in multiple samples, allowing to obtain insight into the potential shared origin of these samples. The main similarity metric reported by StrainGE is the average callable nucleotide identity (ACNI), which is measured as the percentage of positions with strong evidence for the reference allele, comparable to the Average Nucleotide Identity (ANI), taking into account regions of the genome for which variants cannot be called due to the presence of other strains that are very similar or due to insufficient depth of coverage. van Dijk et al. (2022) proposed a threshold for the ACNI of 99.95 % to delineate strains, which was retained in this study.

2.4.4. Phylogenetic tree construction

As a complementary approach to StrainGE, an in-house workflow was developed, based on a pipeline developed by Bogaerts et al. (2024), to perform read mapping, variant calling and construction of SNP-based phylogenetic trees, combining both Illumina and the available ONT sequencing data, which was applied to the three *Bacillus* GMM host strains. This workflow depended on the CFSAN SNP pipeline 2.0.2 (Davis et al., 2015), Bowtie 2.3.4.3, minimap2 2.17, BamTools 2.5.15 (Barnett et al., 2011), Clair3 v0.1-r6 (Zheng et al., 2022), bcftools 1.13 (Danecek et al., 2021), Gubbins 3.1.4 (Croucher et al., 2015), PHASTER (Arndt et al., 2016), BEDtools 2.27.1 (Quinlan & Hall, 2010), and MEGA 10.0.4 (Kumar et al., 2008). The resulting phylogenetic trees were visualized with iTOL (Letunic & Bork, 2024) (main figure) or with FigTree (Rambaut, 2016) (supplementary figures).

2.5. Genotypic detection of AMR and virulence genes

Genotypic AMR detection with KMA (Clausen et al., 2018) on unfiltered short reads was performed as described by Bogaerts et al. (2021), with minor modifications, i.e. only hits with ≥90 % identity and ≥ 90 % target coverage were retained, and instead of the ResFinder database, the National Database of Antibiotic Resistant Organisms (NDARO) (retrieved on 2023-02-05) was used, complemented with an in-house *Bacillus*-specific AMR gene (*catA*, CP023729.1:2725109–2,725,759), which was not present in NDARO. To account for the differences in sample size, the obtained values were normalized by calculating the obtained depth per million unfiltered read pairs. The presence of full-length AMR genes on single raw long reads was assessed with genotypic AMR detection with BLAST+ 2.6.0 according to Bogaerts et al. (2021), with the same modification of the database as described above for the KMA-based gene detection workflow. The unfiltered long-read data, converted to fasta format, was used as query, with the % identity and coverage cutoffs set to 90 % and 95 %, respectively. To investigate whether the open reading frames of genes detected at an identity and/or coverage below 100 % were likely to be structurally intact or not, gene detection with GAMMA 2.1 (Stanton et al., 2022) with default settings was additionally run on the assemblies with the NDARO database.

Genotypic virulence gene detection on unfiltered short-read datasets with KMA was performed as for the AMR genes, with the VFDB-Core as a database (retrieved on 2022-04-20) (Liu et al., 2022) and the % identity and coverage cutoffs set to 90 %.

2.6. Data availability

Raw data and metagenomic assemblies are available in the European Nucleotide Archive under Project accession numbers PRJEB79645 and PRJEB53495.

3. Results

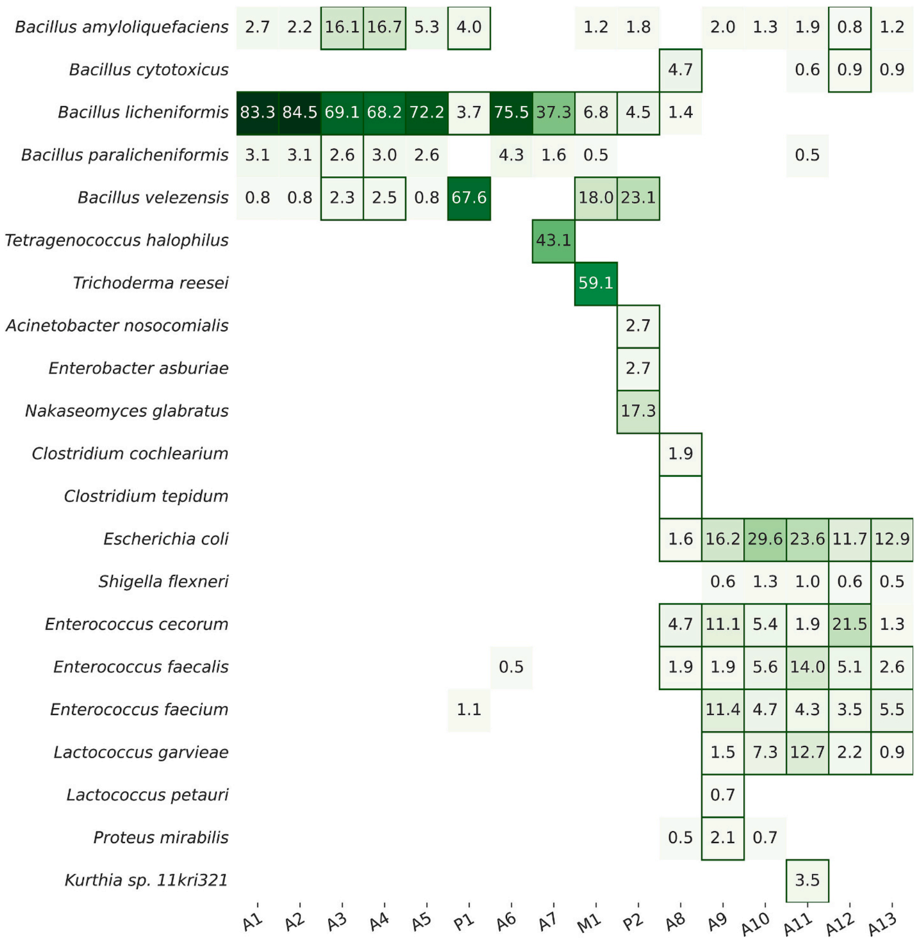
3.1. qPCR demonstrates cross-contamination of food enzyme products with multiple GMM

16 commercial food enzyme (FE) products from 11 different brands were selected from previous studies (D'aes et al., 2021; Fraiture et al., 2022; Fraiture, Deckers, et al., 2020a; Fraiture, Deckers, et al., 2020b; Fraiture, Deckers, et al., 2020c; Fraiture, Gobbo, et al., 2021; Fraiture, Marchesi, et al., 2021), based on their level of GMM contamination observed with qPCR (Table 1, Table S1). For confidentiality reasons, the names of the samples and brands were anonymized with aliases (Table 1) that were used throughout the manuscript to refer to the samples. The FE samples included 13 alpha-amylase FE products (A1-A13), 2 protease products (P1, P2), and 1 mixed product (M1), composed of several enzymes including alpha-amylase and protease (Table 1 qPCR indicated that all samples were contaminated with at least one GMM strain/construct, i.e. GMM protease1, GMM amylase1, and/or GMM amylase2, while none of the samples tested positive for GMM protease2. In general, the dominant contamination reflected the labeled enzyme(s) of the product, although some products appeared to be highly contaminated (Cq < 25) with all three GMM. However, since the qPCR assays target the transgenic construct of the GMM, it is possible that the host strain carrying the construct was not the same in every sample, particularly if the GMM construct resides on a free plasmid. Consequently, qPCR alone did not allow to confirm the presence of the same GMM, i.e. same construct and host strain, and it was therefore not possible to determine whether the contaminations could have shared the same origin based on qPCR results (Table 1). qPCR indicated that all samples were contaminated with at least one GMM strain/construct, i.e. GMM protease1, GMM amylase1, and/or GMM amylase2, while none of the samples tested positive for GMM protease2. In general, the dominant contamination reflected the labeled enzyme(s) of the product, although some products appeared to be highly contaminated (Cq < 25) with all three GMM. However, since the qPCR assays target the transgenic construct of the GMM, it is possible that the host strain carrying the construct was not the same in every sample, particularly if the GMM construct resides on a free plasmid. Consequently, qPCR alone did not allow to confirm the presence of the same GMM, i.e. same construct and host strain, and it was therefore not possible to determine whether the contaminations could have shared the same origin based on qPCR results.

3.2. Metagenomics allows strain-level detection and phylogenomic investigation of relationships of GMM host strains in FE samples

3.2.1. Detection of GMM host species in metagenomic FE samples

Section 2.4 and Fig. 1 provide an outline of the metagenomic workflow and tools used to characterize the samples. Fig. 2 presents the most abundant microbial contaminations detected in the samples at species level (Fig. S1 provides a taxonomic profile at genus level). For all samples, Kraken2 indicated the presence of one or more *Bacillus* spp., which was confirmed by read mapping and/or de novo assembly for samples A1-A8, A12, P1,P2 and M1, but not for samples A9, A10, A11 and A13. The detected species were generally in line with the expected GMM host species based on qPCR (Table 1, Table S1), with *B. licheniformis* (GMM amylase2) and/or *B. amyloliquefaciens* (GMM amylase1) representing the most abundant GMM contaminations in alpha-amylase products (A), while *B. velezensis* (GMM protease1) was the main GMM host species contamination of the protease (P) products. An exception is P2, in which Kraken2 detected *B. velezensis* (GMM protease1) at the highest abundance, in accordance with its product labeling, but in contrast with the qPCR result indicating that GMM amylase1 was more abundant. In the mixed enzyme product M1, all three GMM host species were detected by qPCR, in accordance with its product labelling as well as with the results of Kraken2, although read mapping



	A1	A2	A3	A4	A5	P1	A6	A7	M1	P2	A8	A9	A10	A11	A12	A13
Bacillus amyloliquefaciens	2.7	2.2	16.1	16.7	5.3	4.0			1.2	1.8		2.0	1.3	1.9	0.8	1.2
Bacillus cytotoxicus											4.7			0.6	0.9	0.9
Bacillus licheniformis	83.3	84.5	69.1	68.2	72.2	3.7	75.5	37.3	6.8	4.5	1.4					
Bacillus paralicheniformis	3.1	3.1	2.6	3.0	2.6		4.3	1.6	0.5					0.5		
Bacillus velezensis	0.8	0.8	2.3	2.5	0.8	67.6			18.0	23.1						
Tetragenococcus halophilus								43.1								
Trichoderma reesei									59.1							
Acinetobacter nosocomialis										2.7						
Enterobacter asburiae										2.7						
Nakaseomyces glabratus										17.3						
Clostridium cochlearium											1.9					
Clostridium tepidum																
Escherichia coli											1.6	16.2	29.6	23.6	11.7	12.9
Shigella flexneri												0.6	1.3	1.0	0.6	0.5
Enterococcus cecorum											4.7	11.1	5.4	1.9	21.5	1.3
Enterococcus faecalis							0.5				1.9	1.9	5.6	14.0	5.1	2.6
Enterococcus faecium						1.1						11.4	4.7	4.3	3.5	5.5
Lactococcus garvieae												1.5	7.3	12.7	2.2	0.9
Lactococcus petauri												0.7				
Proteus mirabilis											0.5	2.1	0.7			
Kurthia sp. 11kri321														3.5		

Fig. 2. Taxonomic profile of samples indicating the species detected with Kraken2. Values show the relative abundance obtained with Bracken for species detected with a read abundance of at least 0.5 % (corrected for unclassified reads), and for which detection was confirmed by metagenomic assembly and/or read mapping in at least one sample. Boxed fields indicate that the presence of the species was confirmed in that sample by either read mapping or the presence of a MAG (an empty green lined box represents a MAG that was not detected with Kraken2). The green shading indicates a measure of the relative abundance. The relative abundances per sample do not sum to a hundred because not all reads were classified at species level, and because the presence of a species detected by Kraken2 was not always confirmed. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and/or assembly did not confirm the presence of *B. amyloliquefaciens*.

3.2.2. GMM contaminations are associated with closely related host strains

Fig. 3 shows the StrainGE results for the most relevant *Bacillus* spp. contaminations, i.e. the GMM host, as well as a *B. cytotoxicus* strain that was detected in multiple samples (discussed in section 3.3.3). For the three GMM host strains, the average callable nucleotide identity (ACNI) with the strain in the samples was at least 99.96 % in almost all cases. According to the threshold for the ACNI proposed by van Dijk et al. (2022) to delineate strains (99.95 %), these results indicated that the same GMM host strains were present in the different samples.

In 10 out of 16 samples, the presence of the same GMM amylase1 host strain was confirmed, albeit at a rather low relative abundance, reaching a maximum of 13.42 % for sample A4. A notable exception was P2, for which, although its closest reference was the GMM amylase1 host strain, the ACNI was only 99.93 %, falling below the threshold set to confirm strain identity. In this sample, GMM protease1 (*B. velezensis*) was the dominant GMM contamination, while GMM amylase1 (*B. amyloliquefaciens*) was only present in a minor quantity. Concerning the other two GMM, the same GMM amylase2 host strain was detected in 10 out of 16 samples, with a relative abundance ranging from 0.57 % in A8 to 87.16 % in A6. With the exception of P1, GMM amylase2 was the most abundant GMM host strain contamination in those samples. Seven samples were contaminated with the GMM protease1 host strain, with the relative abundance varying between 0.54 % in A2 and 65.47 % in P1. GMM protease1 was the main GMM host strain contamination of the protease products. In total, seven samples were contaminated with two or even three different GMM strains, including M1, in which both GMM protease1 and GMM amylase2 were found, in line with its mixed enzyme composition. However, the result for M1 is not in agreement with the qPCR assays, which also detected a contamination with GMM amylase1, albeit at a lower level than the contamination with GMM protease1 and GMM amylase2 (Table 1). In several alpha-amylase samples, two distinct alpha-amylase production strains were detected, i.e. GMM amylase1 and GMM amylase2, suggesting that the FE products were composed of mixtures of alpha-amylase extracts obtained with different production strains.

In some samples, the presence of more than one *B. licheniformis* strain was suggested (Table S4), most notably a *B. licheniformis* strain for which viable isolates were obtained from A3 and A5 in previous studies (D'aes et al., 2022; Deckers, Vanneste, et al., 2020). However, the GMM amylase2 host strain was always the most abundant *B. licheniformis* strain, except for P2, for which StrainGE indicated the presence of a *B. licheniformis* strain at 7.2 % relative abundance, while GMM amylase2 was absent from this sample (Table S4). Additionally, in one sample (A11), a low-abundance contamination with a *B. paralicheniformis* strain was detected (Table S4). This contradicts the results of the Kraken2-based taxonomic profile in which the presence of a *B. paralicheniformis* contamination was supported by short-read mapping for sample A3 instead. Finally, five samples were contaminated with *B. cytotoxicus* (see section 3.3.3).

3.2.3. SNP-based phylogenomic analysis supports the close relatedness of GMM host strains in different samples

For all three GMM, the FE metagenomic samples typically clustered together with their respective host strains with high bootstrap support, thereby supporting the close relatedness of the GMM strains in the different samples (Fig. 4, Fig. S2, and Fig. S3). As this approach included a filtering step to retain only regions considered as 'callable' by StrainGE for the respective host strains, the retained region for this analysis did not include the full genome. For this analysis, the retained genome fractions were 55 %, 68 %, and 68 % for GMM amylase2, GMM amylase1, and GMM protease1, respectively, thus taking into account at least half of the genome, which still amounts to several Mbp. In terms of absolute SNP distances (based on the retained genome fractions), the largest distance between the GMM amylase2 host strains in the FE samples was 494 SNPs, while for GMM amylase1, and GMM protease1, the largest absolute SNP distances were 11 and 15, respectively (Table S5, Table S6, Table S7). The relatively large maximal SNP distance between the GMM amylase2 strains is partly due to sample A6, which appeared to differ slightly from the GMM amylase2 strains in other samples. Excluding A6, the remaining maximal SNP distance between the GMM amylase2 host strains was 271. The presence of multiple *B. licheniformis* strains in some of the samples (Table S4) may have inflated the number of SNPs. For instance, in the phylogenetic tree for GMM amylase2 (Fig. 4), P2 clustered distantly from the other FE samples, since no GMM amylase2 contamination was detected in the sample, while another *B. licheniformis* strain was present according to StrainGE (3.2.2).

3.3. Microbial contamination profiles yield further insights in the composition and origin of the samples

3.3.1. Contamination with presumed production organisms

For seven out of the 16 FE samples, the production organism was unknown or described generically as 'bacteria' (Table 1). Of the remaining nine samples, 6 were alpha-amylases and two were proteases, for which the production organism was described as *B. licheniformis* (A5, A6, A7, P2), *B. amyloliquefaciens* (A10), or *B. subtilis* (A12, A9, P1). The ninth sample, M1, constitutes an enzyme mix of protease, cellulase, xylanase, alpha-amylase, and beta-glucanase, and several production organisms were listed, including *B. subtilis*.

Concerning P2, a *B. licheniformis* strain was indeed detected, but the major *Bacillus* contaminant was *B. velezensis* (GMM protease1), in line with its enzymatic composition. In A12 and A9, *B. amyloliquefaciens* (GMM amylase1) was present instead of *B. subtilis*, and was presumably the producer organism, while in P1 *B. velezensis* (GMM protease1) was the most likely producer organism, although the GMM amylase1 and amylase2 were also detected. Conversely, multiple alpha-amylase samples, i.e., A1, A2, A3, and A4, were cross-contaminated with GMM protease1 strains. Finally, in M1, the mixed enzyme sample, *B. licheniformis* (GMM amylase2) and *B. velezensis* (GMM protease1) were detected. Together, these results indicate that the labeling of the production organism as *B. subtilis* for A12, A9, P1, and M1, and the labeling as *B. licheniformis* for P2, is correct when considered at a higher level, since the detected species are part of the *B. subtilis group*, to which e.g. *B. licheniformis*, *B. amyloliquefaciens*, and *B. velezensis* belong.

In addition to GMM amylase2 and GMM protease1, the listed production organism *Trichoderma reesei* was detected in M1. It is noteworthy that the contamination with *T. reesei* was so extensive that a nearly complete assembly of the 43 Mbp genome could be recovered, showing nearly 100 % nucleotide identity to strain QM6a (Table S3). Hyperproducing mutants derived from this strain by random mutagenesis are one of the most widely used microbial producers of cellulase (Le Crom et al., 2009).

3.3.2. Contamination with micro-organisms unrelated to the fermentation products

In seven samples (A1, A2, A3, A4, A5, P1, A6), the microbial contamination was primarily classified as *Bacillus* (from 96 to 99 % according to Bracken, cfr. Fig. S1), and strongly associated with the GMM contamination, with little or no presence of other microbial species (Fig. 2). Several other contaminations were however observed with Kraken2 for other evaluated FE samples. In these samples, *Bacillus* often constituted only a minor fraction of the contamination, while other species dominated the microbial contamination profile (Fig. 2).

The contamination profile of A7 was made up of roughly equal fractions *Bacillus* and *Tetragenococcus halophilus*, a halophilic lactic acid bacterium commonly employed in the fermentation processes of soy sauce, miso, fish sauce and salted anchovies (Justé et al., 2014). The most notable contamination in the P2 sample, next to *Bacillus*, was *Nakaseomyces glabratus*, previously known as *Candida glabrata*, which is the dominant microbial producer for the chemical compound pyruvic acid (Luo et al., 2020). It is also an opportunistic fungal pathogen that currently ranks as the second most common cause of candidiasis (Carreté et al., 2019). Additionally, P2 was contaminated with *Acinetobacter nosocomialis*, known as a Gram-negative opportunistic pathogen (Knight et al., 2018), and *Enterobacter asburiae,* which is ubiquitous, with studies reporting a range of potentially beneficial as well as detrimental effects on plants, animals, and humans (Francis et al., 2020; Horinouchi et al., 2022; Oh et al., 2018; Xue et al., 2021). A8 was contaminated with a diverse range of micro-organisms, including *Bacillus cytotoxicus*, *Escherichia coli*, *Enterococcus cecorum*, *Enterococcus faecalis*, and several *Clostridium* spp., a genus with important applications in diverse domains of industrial biotechnology. Although Bracken classified 53 % of the entire sample as *Clostridium* at genus level (Fig. S1), the read-based and assembly-based taxonomic classifications were not in agreement concerning the species, which illustrates its status as a 'problematic' genus for classification (Cruz-Morales et al., 2019). The metagenomic assembly (Table S3) indicated the presence of the thermophilic species *C. tepidum*, while read-mapping supported the Kraken2-detected species *C. cochlearium.*

The five remaining samples (A9, A10, A11, A12, A13) presented a similar microbial contamination profile, in which the dominant contaminations were *Enterococcus* spp., *E. coli/Shigella flexneri,* and *Lactococcus garvieae*. Furthermore, according to StrainGE (Fig. 3), four out of five samples were contaminated with a low-abundant presence of *B. cytotoxicus*, which belongs to the *B. cereus* group and is a potentially pathogenic species. Besides these common contaminations, A9 was also contaminated with *Lactococcus petauri* and *Proteus mirabilis,* while A11 contained *Kurthia* sp. 11kri321. *E. coli* is known for its use in biotechnological processes, but contains pathogenic members as well, while *S. flexneri* is predominantly known to be pathogenic. *E. coli* and *S. flexneri* are closely related, which may explain the read-based taxonomic classification supporting the detection of *E. coli*, while the assembly-based classification supported the detection of *S. flexneri* (Jin et al., 2002). L. *garvieae* and *L. petauri* are the etiological agents of *Lactococcosis*, an emerging disease affecting many fish species, leading to economic losses (Vendrell et al., 2006). All samples were contaminated with three *Enterococcus* species, i.e. *E. faecalis*, *E. faecium*, and/or *E. cecorum*, with the exception of sample A13, in which the presence of *E. cecorum* was indicated by Kraken2, but not confirmed by metagenomic assembly nor read mapping. The species *E. faecium* and *E. faecalis* encompass both clinical strains as well as non-pathogenic strains that are commonly used in microbial fermentation processes or in probiotic applications (Franz & Holzapfel, 2004), while *E. cecorum* is mainly associated with poultry infections (Jung et al., 2018).

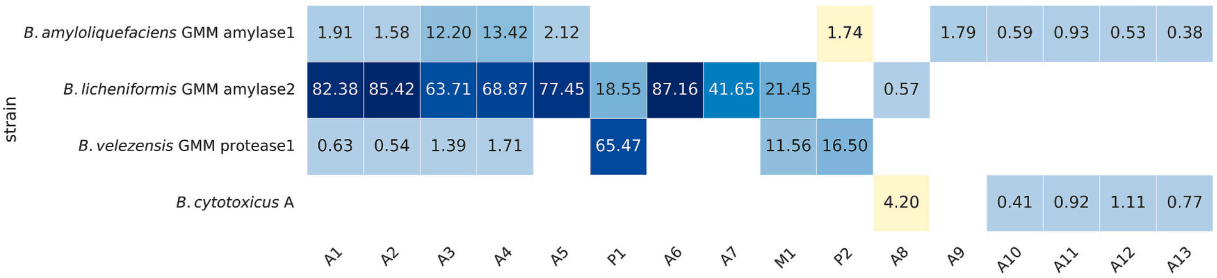


Fig. 3. Overview of strain-deconvolution results for the most relevant detected *Bacillus* strains with StrainGE. The values represent the relative abundance of the strain in the sample as detected by StrainGST. For values highlighted in blue the average callable nucleotide identity (ACNI) of the strain in the sample to the strain in the database was at least 99.95 % (or for *B. cytotoxicus* to the strain in the sample with the highest relative reported relative abundance for this strain), indicating that they are the same strains. The blue shading reflects the relative abundance of the strain in the sample. Conversely, if the ACNI was below 99.95 %, the value is highlighted in yellow. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

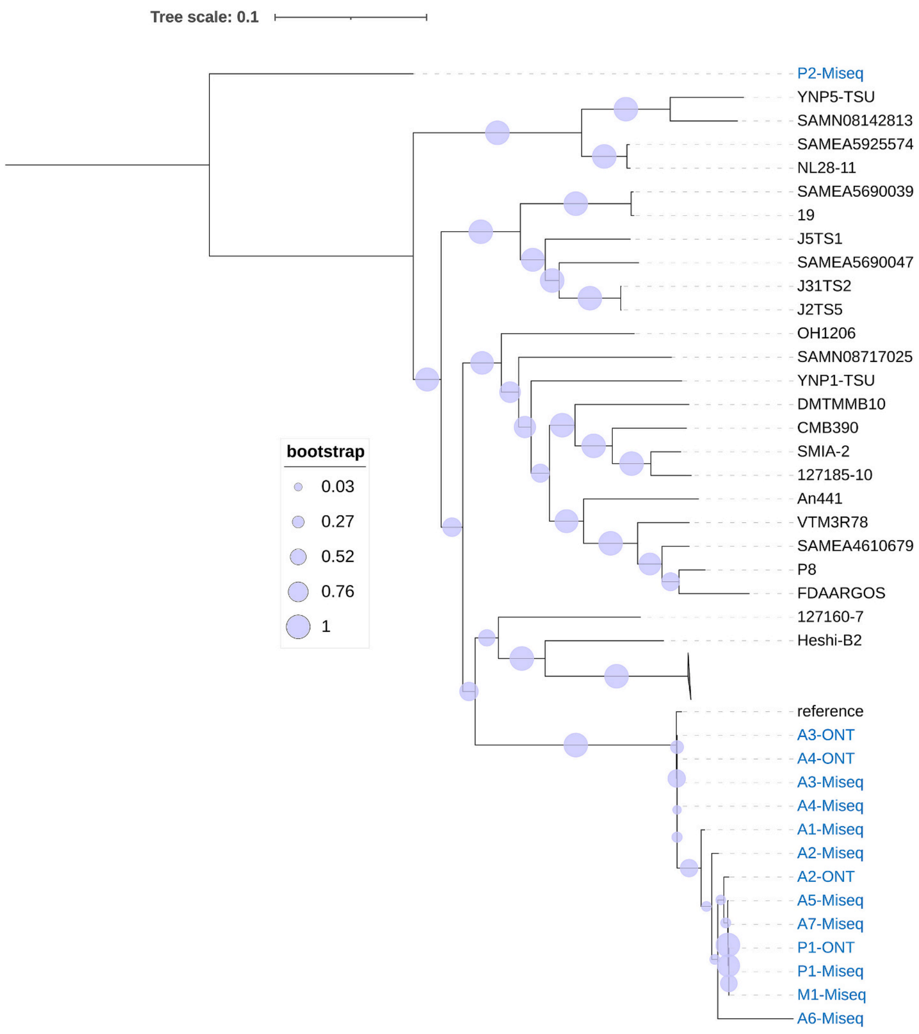


Fig. 4. SNP-phylogenetic tree for GMM amylase2. The scale bar is expressed as average substitutions per site in the SNP matrix. Node values represent bootstrap support values (as decimals). Blue colored names correspond to samples included in this study (Table 1). Some samples were sequenced with both Illumina (Miseq) and ONT technology, as denoted with the suffix. The SNP-phylogenetic trees for GMM amylase1 and GMM protease1 are presented in Fig. S2 and Fig. S3. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.3.3. *Five samples likely share common Enterococcus and B. cytotoxicus strains*

For samples A9, A10, A11, A12, and A13 a similar microbial contamination profile was observed (Fig. 2, Figs. 3, 3.3.2). Additionally, sample A8 also exhibited a similar contamination profile with several *Enterococcus* spp., but also some differences such as the absence of *E. faecium* and L. *garvieae*, and the presence of *B. cytotoxicus* at a higher relative abundance. To investigate a potential shared origin of these samples, the *Enterococcus* and *B. cytotoxicus* strains in the six samples were analyzed with StrainGE. This indicated that for both species, the strains in the samples were different from the strains in the underlying reference database. To highlight that they did not match any of the database strains, the strains were designated provisional names, e.g. *E. cecorum* A (Fig. 5). The results suggested that multiple *Enterococcus* strains were shared among samples A9, A10, A11, A12, and A13, while this was not observed for sample A8. Likewise, the results indicated that A10, A11, A12 and A13 were contaminated with the same *B. cytotoxicus* strain, while in A8 it appeared to be another strain (Fig. 3).

The highly similar overall microbial contamination profile, together with the suspicion of a shared origin of multiple *Enterococcus* strains and the *B. cytotoxicus* strain, supported a common origin of samples A9, A10, A11, A12, and A13. Furthermore, these results indicated that the *B. cytotoxicus* and *Enterococcus* strains, with exception of *E. faecalis* C, in A8 differed from those in the other five samples. Together with the additional conspicuous differences in its microbial contamination profile (see above), this suggested the contaminations in A8 to be unrelated to these five other samples.

3.4. Presence of AMR genes in the samples is strongly associated with GMM contamination

Finally, to further assess the potential public health risk associated with the microbial contaminations in the FE products, the presence of AMR and virulence genes in the samples was investigated. The AMR profile of the samples (Fig. 6), clearly highlighted that the GMM contamination was the primary source of AMR genes in the samples. *aadD1* and *bleO* originate from the pUB110 vector, which was used to generate the transgenic constructs of GMM amylase1 and GMM protease1. *catA* is part of the *catA-amyS* transgenic construct that was chromosomally integrated in GMM amylase2, while *blaP* is a naturally occurring gene from the GMM amylase2 host strain *B. licheniformis*. Moreover, in samples for which long-read data was available, full-length copies of these genes were detected on single raw long reads. In several samples, a variety of other AMR genes was detected, in line with their varied microbial contamination profiles, although most of these genes were detected at much lower depths compared to the GMM-associated AMR genes. Likewise, the virulence gene profile of the samples (Fig. S4) indicated the presence of a wide variety of virulence genes in certain samples, specifically those with a more varied microbial contamination profile, although in the majority of samples, not a single virulence gene was detected.

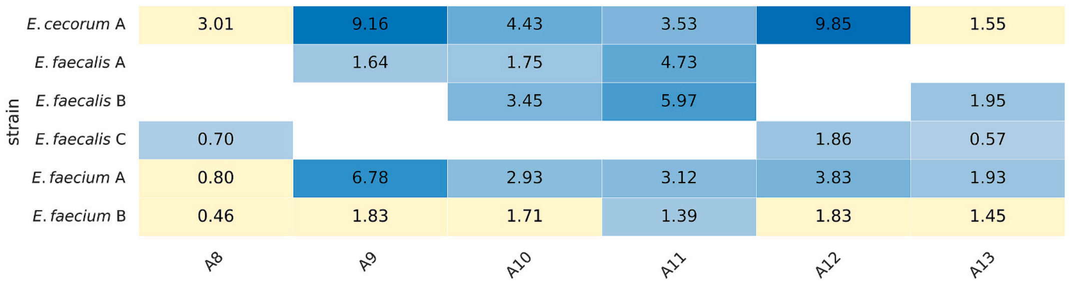
4. Discussion

Microbially produced food enzymes, such as alpha-amylase and protease, are widely used in the food industry, especially in bakeries where they are e.g. added to the dough to improve bread quality, but also for starch liquefaction, in breweries, as digestive aid, etc. (Raveendran et al., 2018). The use of GM strains for food enzyme (FE) production is widespread, frequently leading to product contamination with recombinant DNA or living GMM. Recent pilot surveillance studies indicated that a significant fraction of the FE products, collected from the EU market, were contaminated with recombinant DNA of one or multiple GMM (Deckers et al., 2022; Fraiture et al., 2024). Fraiture et al. (2024), employing newly developed qPCR methods, found that 55 % (22/40) of the FE samples was GMM-contaminated, of which 59 % (13/22) at a high contamination level (Cq < 25). Similar contamination issues are likely also present outside of the EU, considering the global nature of the food enzyme market. However, to our knowledge, no similar studies concerning GMM contamination of microbial fermentation products outside of the EU are available.

Currently, qPCR constitutes the state-of-the-art approach to screen samples for the presence of GMM contaminations (Barbau-piednoir et al., 2015; Fraiture et al., 2022; Fraiture et al., 2024; Fraiture, Bogaerts, et al., 2020; Fraiture, Deckers, et al., 2020b; Fraiture, Deckers, et al., 2020c; Fraiture, Gobbo, et al., 2021; Fraiture, Marchesi, et al., 2021). Because the qPCR assays typically target the GMM construct, qPCR alone cannot always confirm the presence of the same GMM in different samples, e.g. if the GMM construct resides on an episomal plasmid as different host strains could potentially harbor the same plasmid construct. Furthermore, development of a qPCR assay requires prior knowledge concerning the intended target, and only allows to detect the targeted species, strain or construct. Separate assays must also be designed and validated for each GMM construct, requiring a substantial amount of lab work Similarly, performing the different assays on each sample implies a significant amount of hands-on labor.

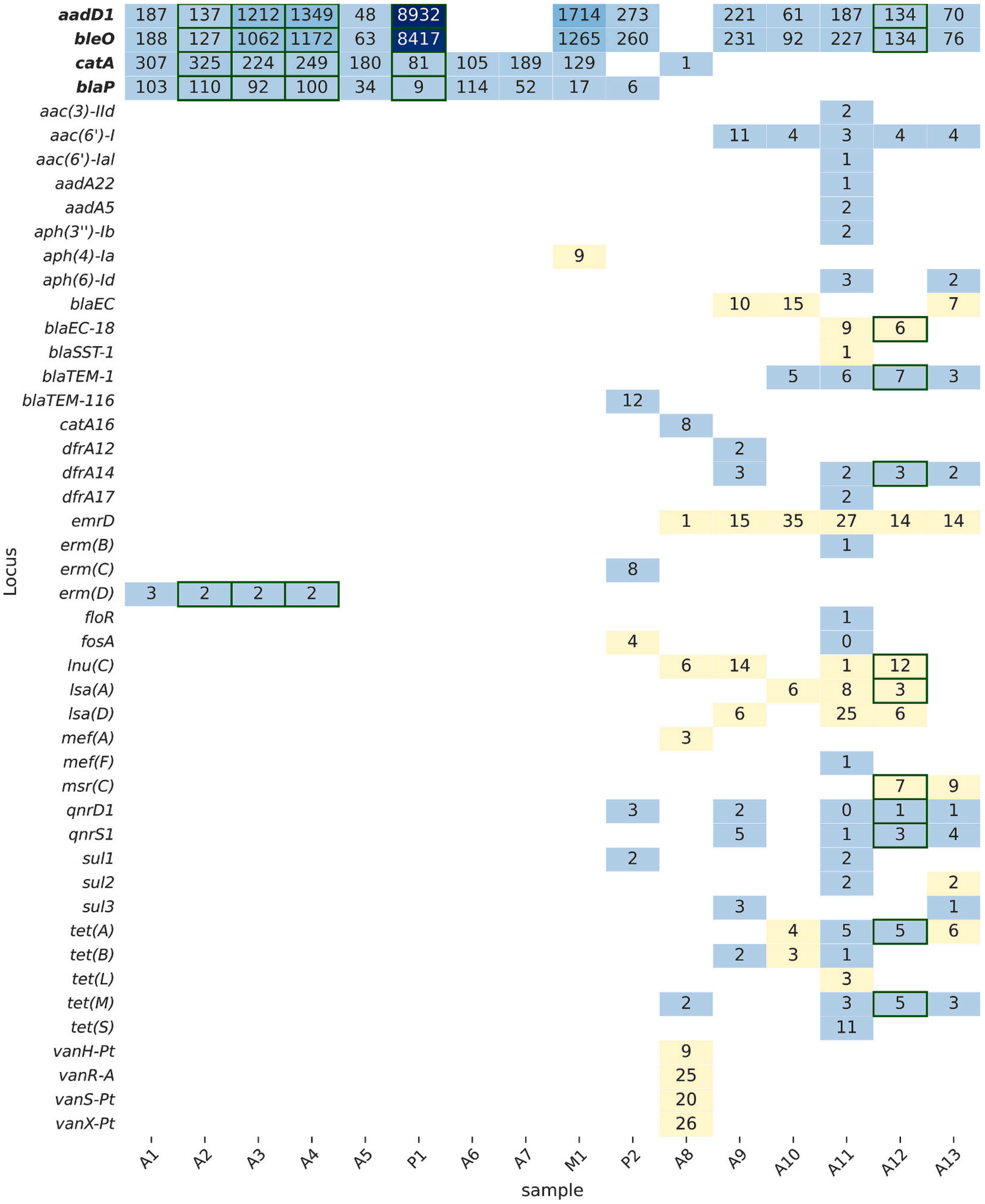
Metagenomics constitutes an attractive alternative approach because of several advantages. With a shotgun metagenomics approach, detection and characterization of the GMM constructs and their host strains becomes feasible, although it does not allow to link episomal GMM constructs to their host directly (Buytaers et al., 2021; D'aes et al., 2022). It can also allow detection and characterization of GMM constructs in species that cannot be cultured for subsequent isolate WGS. Since metagenomics-based screening is not restricted to the envisaged species or strains, it can even reveal the presence of previously uncharacterized GMM or other unexpected contaminations (D'aes et al., 2022). Lastly, metagenomic data allows to screen simultaneously for all GMM sequences that have been previously identified and characterized, which could account for a substantial decrease in the amount of hands-on work required. However, the data analysis and results interpretation currently still requires substantial bioinformatics expertise.

Strain-level deconvolution and phylogenomic investigation in metagenomic samples to perform source tracing of contaminations, including GMM, remains a significant challenge. In a previous study, source tracing of GMM was demonstrated for isolates of GMM protease1 (D'aes et al., 2021). However, viable isolates of GMM amylase1 and GMM amylase2 could however not be obtained, possibly due to knockouts in several sporulation genes, precluding the use of a WGS-based



strain	A8	A9	A10	A11	A12	A13
E. cecorum A	3.01	9.16	4.43	3.53	9.85	1.55
E. faecalis A		1.64	1.75	4.73		
E. faecalis B			3.45	5.97		1.95
E. faecalis C	0.70				1.86	0.57
E. faecium A	0.80	6.78	2.93	3.12	3.83	1.93
E. faecium B	0.46	1.83	1.71	1.39	1.83	1.45

Fig. 5. Overview of strain-deconvolution results for detected *Enterococcus* strains with StrainGE in samples A8, A9, A10, A11, A12, and A13. The values represent the relative abundance of the strain in the sample as detected by StrainGST. The strains were labeled with A, B, and C to distinguish them from each other as well as from the strains in the reference database. Values highlighted in blue indicate that the average callable nucleotide identity (ACNI) of the strain in the sample was at least 99.95 % to the reference sample (i.e. the sample with the highest reported relative abundance for this strain), indicating that they are the same strains. The blue shading reflects the relative abundance of the strain in the sample. Conversely, if the ACNI was below 99.95 %, the value is highlighted in yellow. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



Locus	A1	A2	A3	A4	A5	P1	A6	A7	M1	P2	A8	A9	A10	A11	A12	A13
aadD1	187	137	1212	1349	48	8932			1714	273		221	61	187	134	70
bleO	188	127	1062	1172	63	8417			1265	260		231	92	227	134	76
catA	307	325	224	249	180	81	105	189	129		1					
blaP	103	110	92	100	34	9	114	52	17	6						
aac(3)-IId														2		
aac(6')-I												11	4	3	4	4
aac(6')-Ial														1		
aadA22														1		
aadA5														2		
aph(3'')-Ib														2		
aph(4)-Ia									9							
aph(6)-Id														3		2
blaEC												10	15			7
blaEC-18														9	6	
blaSST-1														1		
blaTEM-1													5	6	7	3
blaTEM-116										12						
catA16											8					
dfrA12												2				
dfrA14												3		2	3	2
dfrA17														2		
emrD											1	15	35	27	14	14
erm(B)														1		
erm(C)										8						
erm(D)	3	2	2	2												
floR														1		
fosA										4				0		
lnu(C)											6	14		1	12	
lsa(A)													6	8	3	
lsa(D)												6		25	6	
mef(A)											3					
mef(F)														1		
msr(C)															7	9
qnrD1										3		2		0	1	1
qnrS1												5		1	3	4
sul1										2				2		
sul2														2		2
sul3												3				1
tet(A)													4	5	5	6
tet(B)												2	3	1		
tet(L)														3		
tet(M)											2			3	5	3
tet(S)														11		
vanH-Pt											9					
vanR-A											25					
vanS-Pt											20					
vanX-Pt											26					

Fig. 6. Overview of AMR gene load in the samples. The values represent the relative abundance, normalized per million read pairs, of the gene in the short-read samples. Only genes that were considered potentially functional with GAMMA are shown (i.e., not marked as 'truncated' or 'contig edge'), with the green shading reflecting the relative abundance of the gene in the sample. Genes with mutations compared to the database reference are highlighted in yellow. Genes associated with the GMM contaminations are shown in bold. Genes for which full-length copies were detected on single raw long reads (only available for samples A2, A3, A4, P1 and A12) are boxed. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

approach for source tracing of these GMM (D'aes et al., 2022). Although findings of the latter study hinted at the close similarity of the GMM amylase1 and GMM amylase2 strains in different FE samples, the employed methodology did not allow performing source tracing on the host strains of the GMM contaminations. A considerable range of short-read based strain deconvolution tools exists, including both reference-based tools relying on a database as well as reference-free de novo methods. Although long-read sequencing approaches could potentially improve accuracy, long-read tools tailored to strain deconvolution and phylogenomics in metagenomics samples remain scarce. Based on a preliminary investigation (unpublished results), StrainGE was selected for this study as it appeared to provide robust results (Lindstedt et al., 2022; Salamzade et al., 2023; van Dijk et al., 2022). Moreover, although it is a reference-based method, it is capable of identifying and comparing strains that are not represented in the provided database, representing a substantial added value, especially with regard to the detection of uncharacterized GMM contaminations. However, StrainGE is tailored to processing paired Illumina data, and does not allow to perform a SNP-based phylogenetic analysis of the detected strains. To the best of our knowledge, currently no bioinformatics tools are available that allow strain-aware SNP-based reconstruction of phylogenetic trees using both long and short-read metagenomic data. Therefore, we developed a custom SNP-based tree reconstruction workflow, compatible with both short and long-read data, in which we exploited the ability of StrainGE to define genomic regions with sufficient intra-strain variability to allow strain deconvolution. This workflow was run on several target strains of interest, allowing to combine and compare both short- and long-read data.

Overall, the results obtained with StrainGE were in agreement with qPCR results, as well as with previous findings indicating that this tool produces reliable results, and can be of added value to screen microbial fermentation products for the presence of GMM and other microbial contaminants at strain level. Additionally, StrainGE allowed to compare the *Bacillus* GMM strains in different samples, supporting the presence of the same GMM host strains in multiple samples (Fig. 3). Similarly, StrainGE indicated that several strains of *Enterococcus* contaminations were shared among five different samples (Fig. 5), allowing to derive suspicion concerning a common origin of these samples. The SNP-based phylogenomic analysis confirmed that the *Bacilli* GMM host strains were very closely related (Fig. 4, Fig. S2, Fig. S3), differing only maximum 11 and 15 SNPs for amylase 1 and protease1, respectively, whereas the amylase2 host strains displayed larger variation with up to 271 SNPs. The results from the short- and long-read data for a given sample were in good agreement, indicating that this analysis approach allows to combine data obtained with different sequencing platforms. The results of the SNP-based phylogenomic analysis were generally in line with those of StrainGE, although in some cases, the insights provided by StrainGE were needed to correctly interpret the phylogenetic trees. For instance, a *B. licheniformis* strain distinct from GMM amylase2 was detected in P2 by StrainGE, resulting in P2 clustering distantly from the other FE samples in the GMM amylase2 tree. This illustrates a limitation of this approach, as accurate placement in the tree requires that the targeted strain, e.g. in this case GMM amylase2, is the dominant strain of the species in the sample. The results of the SNP-based tree reconstruction may hence be inaccurate if multiple strains of the same species are present in the sample. Consequently, StrainGE and the SNP-based phylogenomic analysis supported that in most samples, the GMM amylase1, GMM amylase2, and GMM protease1 host strains were derived from the same parental GMM strain. To our knowledge, this is the first time that source tracing of GMM contaminations is achieved based on a cultivation-independent, metagenomic approach. Insight into the contamination source is of crucial relevance to allow the competent authorities to take appropriate actions.

According to the qPCR results, twelve out of the sixteen samples included in this study were highly contaminated (Cq < 25) with at least one GMM (Table 1). This study additionally demonstrated the presence of single DNA molecules carrying complete copies for all GMM-derived AMR genes that were detected in the samples for which long-read data was available. According to the copy number estimates obtained with the short-read data, these AMR genes were present at a high abundance (Fig. 6). Therefore, there exists a potential risk for horizontal gene transfer of these genes, leading to spreading to other microorganisms, even if the contamination is no longer associated with a living GMM (Arnold et al., 2022). It is currently not clear to what extent such a transfer of AMR genes from a GMM-contaminated product to other microorganisms is feasible, or even possible. More research is needed to investigate this, and will yield valuable results concerning the evaluation of the risk involved in such contaminations.

Overall, most of the additional microbial contaminations on top of the Bacilli in the samples hinted at an origin in the production environment. Some of the detected species were associated with microbial fermentations, such as *T. halophilus*, *Clostridium*, and *T. reesei*, while other species are well known as causal agents of diseases common in livestock, e.g. *E. cecorum*, *L. garvieae*, and *P. larvae*. The origins of these contaminations in the food enzyme products are uncertain, and can only be speculated upon. Some of them could have been present as contaminants of the substrate for the fermentation, while in other cases cross-contaminations with producer strains from other fermentation processes taking place in the same reactor or managed by the same operator, may have occurred. Accidental contaminations with naturally occurring strains is also a possibility during the entire production process of the food enzymes. Based on comparing the microbial contamination profiles obtained with Kraken2 and StrainGE, a shared origin was suggested for at least five samples. Although the samples represent at least three different brands (the brand of sample A13 is unknown), this indicates that they likely originate from the same production facility. Irrespective of the source(s) of the contaminations, their presence signals issues with the imposed sanitation measures at the production facility. However, it is not possible to assess the public health risk associated with these contaminations without further work to determine if some of these strains are viable and/or potentially pathogenic. Although the results for the 16 samples analyzed in this study cannot be extrapolated to draw conclusions regarding FE products or microbial fermentation products in general, the metagenomic approach developed in this study could be applied within a wider scope to characterize the microbial contamination profile of other types of products, in particular with (genetically modified) microbial producer strains.

5. Conclusion

In this study, GMM contaminations in a range of commercial food enzyme products from different brands were characterized and compared with a metagenomic approach without the need for microbial isolation. The results highlighted the potential of metagenomics to investigate unculturable contaminations, taking advantage of the untargeted nature of metagenomics to gain insight into the microbial composition and origin of the samples, allowing source tracing of GM strains directly on metagenomic data. Additionally, the results showcased the added value of long-read sequencing to detect the presence of full-length copies of AMR genes in the samples. In this case study, most of the AMR gene load detected in the samples originated from the GMM contaminations, highlighting the potential public health risk associated with such contaminations in products destined for the food and feed industry.

Funding

This research was funded by the Transversal activities in Applied Genomics Service from Sciensano.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

CRediT authorship contribution statement

Jolien D'aes: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Marie-Alice Fraiture:** Writing – review & editing, Writing – original draft, Validation, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Bert Bogaerts:** Writing – review & editing, Software, Resources, Methodology. **Yari Van Laere:** Writing – review & editing, Software, Resources, Methodology. **Sigrid C.J. De Keersmaecker:** Writing – review & editing, Resources, Methodology. **Nancy H.C. Roosens:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Kevin Vanneste:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Conceptualization.

Declaration of competing interest

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

Acknowledgements

We thank the technicians of the Transversal Activities in Applied Genomics Service for performing the Illumina next-generation sequencing runs.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fochms.2024.100236.

Data availability

Data will be made available on request.

References

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., & Wishart, D. S. (2016). PHASTER: A better, faster version of the PHAST phage search tool. *Nucleic Acids Research, 44*(W1), W16–W21. https://doi.org/10.1093/nar/gkw387

Arnold, B. J., Huang, I. T., & Hanage, W. P. (2022). Horizontal gene transfer and adaptive evolution in bacteria. *Nature Reviews Microbiology, 20*(4), 206–218. https://doi.org/10.1038/s41579-021-00650-4

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., … Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology, 19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Barbau-piednoir, E., De Keersmaecker, S. C. J., Delvoye, M., Gau, C., Philipp, P., & Roosens, N. H. C. (2015). Use of next generation sequencing data to develop a qPCR method for specific detection of EU-unauthorized genetically modified *Bacillus subtilis* overproducing riboflavin. *BMC Biotechnology, 15(1), Article 103*. https://doi.org/10.1186/s12896-015-0216-y

Barnett, D. W., Garrison, E. K., Quinlan, A. R., & Strïmberg, M. P., & Marth, G. T.. (2011). Bamtools: A C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics, 27*(12), 1691–1692. https://doi.org/10.1093/bioinformatics/btr174

Bertrand, D., Shaw, J., Kalathiyappan, M., Ng, A. H. Q., Kumar, M. S., Li, C., … Nagarajan, N. (2019). Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nature Biotechnology, 37*(8), 937–944. https://doi.org/10.1038/s41587-019-0191-2

Bogaerts, B., Nouws, S., Verhaegen, B., Denayer, S., Van Braekel, J., Winand, R., … Vanneste, K. (2021). Validation strategy of a bioinformatics whole genome sequencing workflow for Shiga toxin-producing *Escherichia coli* using a reference collection extensively characterized with conventional methods. Microbial. *Genomics, 7(3), Article 000531*. https://doi.org/10.1099/mgen.0.000531

Bogaerts, B., Van den Bossche, A., Verhaegen, B., Delbrassinne, L., Mattheus, W., Nouws, S., … Vanneste, K. (2024). Closing the gap: Oxford Nanopore technologies R10 sequencing allows comparable results to Illumina sequencing for SNP-based outbreak investigation of bacterial pathogens. *Journal of Clinical Microbiology, 62*(5). https://doi.org/10.1128/jcm.01576-23

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bushnell, B., Rood, J., & Singer, E. (2017). *BBTools software package*. PLOS ONE.

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

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics, 10*, 421. https://doi.org/10.1186/1471-2105-10-421

Carreté, L., Ksiezopolska, E., Gómez-Molero, E., Angoulvant, A., Bader, O., Fairhead, C., & Gabaldón, T. (2019). Genome comparisons of *Candida glabrata* serial clinical isolates reveal patterns of genetic variation in infecting clonal populations. *Frontiers in Microbiology, 10*. https://doi.org/10.3389/fmicb.2019.00112

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. *Bioinformatics, 36*(6), 1925–1927. https://doi.org/10.1093/bioinformatics/btz848

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

Croucher, N. J., Page, A. J., Connor, T. R., Delaney, A. J., Keane, J. A., Bentley, S. D., … Harris, S. R. (2015). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research, 43(3), Article 15*. https://doi.org/10.1093/nar/gku1196

Cruz-Morales, P., Orellana, C. A., Moutafis, G., Moonen, G., Rincon, G., Nielsen, L. K., … Bapteste, E. (2019). Revisiting the evolution and taxonomy of Clostridia, a phylogenomic update. *Genome Biology and Evolution, 11*(7), 2035–2044. https://doi.org/10.1093/gbe/evz096

D'aes, J., Fraiture, M.-A., Bogaerts, B., De Keersmaecker, S. C. J., Roosens, N. H. C., & Vanneste, K. (2021). Characterization of genetically modified microorganisms using short-and long-read whole-genome sequencing reveals contaminations of related origin in multiple commercial food enzyme products. *Foods, 10(11), Article 2637*. https://doi.org/10.3390/foods10112637

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), Article 1971. https://doi.org/10.3390/life12121971

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., … Davies, R. M. (2021). Twelve years of SAMtools and BCFtools. *GigaScience, 10(2), Article giab008*. https://doi.org/10.1093/gigascience/giab008

Davis, S., Pettengill, J. B., Luo, Y., Payne, J., Shpuntoff, A., Rand, H., & Strain, E. (2015). CFSAN SNP pipeline: An automated method for constructing snp matrices fromnext-generation sequence data. PeerJ computer. *Science, 2015(8), Article 1*, 20. https://doi.org/10.7717/peerj-cs.20

Deckers, M., De Loose, M., Papazova, N., Deforce, D., Fraiture, M.-A., & Roosens, N. H. C. (2022). First monitoring for unauthorized genetically modified bacteria in food enzymes from the food market. *Food Control, 135, Article 108665*. https://doi.org/10.1016/j.foodcont.2021.108665

Deckers, M., Deforce, D., Fraiture, M.-A., & Roosens, N. H. C. (2020). Genetically modified micro-organisms for industrial food enzyme production: An overview. *Foods, 9(3), Article 326*. https://doi.org/10.3390/foods9030326

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.2019.1254

van Dijk, L. R., Walker, B. J., Straub, T. J., Worby, C. J., Grote, A., Schreiber, H. L., … Earl, A. M. (2022). StrainGE: A toolkit to track and characterize low-abundance strains in complex microbial communities. *Genome Biology, 23*(1), Article 74. https://doi.org/10.1186/s13059-022-02630-0

Fan, B., Blom, J., Klenk, H. P., & Borriss, R. (2017). *Bacillus amyloliquefaciens*, *Bacillus velezensis*, and *Bacillus siamensis* form an "operational group *B. Amyloliquefaciens*" within the *B. Subtilis* species complex. *Frontiers in Microbiology, 8*. https://doi.org/10.3389/fmicb.2017.00022

Florez-Cuadrado, D., Moreno, M. A., Ugarte-Ruíz, M., & Domínguez, L. (2018). Antimicrobial resistance in the food chain in the European Union. *Advances in Food and Nutrition Research, 86*, 115–136. https://doi.org/10.1016/bs.afnr.2018.04.004

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), 1–12. https://doi.org/10.1038/s41598-020-63987-5

Fraiture, M.-A., Deckers, M., Papazova, N., & Roosens, N. H. C. (2020a). 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., Deckers, M., Papazova, N., & Roosens, N. H. C. (2020b). Detection strategy targeting a chloramphenicol resistance gene from genetically modified bacteria in food and feed products. *Food Control, 108, Article 106873*. https://doi.org/10.1016/j.foodcont.2019.106873

Fraiture, M.-A., Deckers, M., Papazova, N., & Roosens, N. H. C. (2020c). Strategy to detect genetically modified bacteria carrying tetracycline resistance gene in fermentation products. *Food Analytical Methods, 13*(10), 1929–1937. https://doi.org/10.1007/s12161-020-01803-6

Fraiture, M.-A., Gobbo, A., Guillitte, C., Marchesi, U., Verginelli, D., De Greve, J., … Roosens, N. H. C. (2024). Pilot market surveillance of GMM contaminations in alpha-amylase food enzyme products: A detection strategy strengthened by a newly developed qPCR method targeting a GM *Bacillus licheniformis* producing alpha-amylase. *Food Chemistry: Molecular Sciences, 8, Article 100186*. https://doi.org/10.1016/j.fochms.2023.100186

Fraiture, M.-A., Gobbo, A., Marchesi, U., Verginelli, D., Papazova, N., & Roosens, N. H. C. (2021). Development of a real-time PCR marker targeting a new unauthorized genetically modified microorganism producing protease identified by DNA walking. *International Journal of Food Microbiology, 354, Article 109330*. https://doi.org/10.1016/j.ijfoodmicro.2021.109330

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), Article 78*. https://doi.org/10.3390/fermentation8020078

Fraiture, M.-A., Marchesi, U., Verginelli, D., Papazova, N., & Roosens, N. H. C. (2021). Development of a real-time PCR method targeting an unauthorized genetically modified microorganism producing alpha-amylase. *Food Analytical Methods, 14*, 2211–2220. https://doi.org/10.1007/s12161-021-02044-x

Fraiture, M.-A., Papazova, N., & Roosens, N. H. C. (2021). DNA walking strategy to identify unauthorized genetically modified bacteria in microbial fermentation products. *International Journal of Food Microbiology, 337, Article 108913*. https://doi.org/10.1016/j.ijfoodmicro.2020.108913

Francis, M. J., Chin, J., Lomiguen, C. M., & Glaser, A. (2020). Cotton fever resulting in *Enterobacter asburiae* endocarditis. *IDCases, 19, Article e00688*. https://doi.org/10.1016/j.idcr.2019.e00688

Franz, C. M. A. P., & Holzapfel, W. H. (2004). The genus *Enterococcus*: Biotechnological and safety issues. In S. Salminen, & A. von Wright (Eds.), *Lactic acid Bacteria* (pp. 199–248). Microbiological and Functional Aspects, Third Edition: Revised and Expanded. https://doi.org/10.1201/9780824752033.

Graham, A. E., & Ledesma-Amaro, R. (2023). The microbial food revolution. Nature. *Communications, 14*(1), 2231. https://doi.org/10.1038/s41467-023-37891-1

Horinouchi, N., Shiota, S., Takakura, T., Yoshida, A., Kikuchi, K., Nishizono, A., & Miyazaki, E. (2022). Bacteremia caused by *Enterobacter asburiae* misidentified biochemically as *Cronobacter sakazakii* and accurately identified by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry: A case report. *Journal of Medical Case Reports, 16(1), Article 19*. https://doi.org/10.1186/s13256-021-03241-2

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics, 11, Article 119*. https://doi.org/10.1186/1471-2105-11-119

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

Jin, Q., Yuan, Z., Xu, J., Wang, Y., Shen, Y., Lu, W., Wang, J., Liu, H., Yang, J., Yang, F., Zhang, X., Zhang, J., Yang, G., Wu, H., Qu, D., Dong, J., Sun, L., Xue, Y., Zhao, A., & Yu, J. (2002). Genome sequence of *Shigella flexneri* 2a: Insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Research, 30*(20), 4432–4441. https://doi.org/10.1093/nar/gkf566

Jung, A., Chen, L. R., Suyemoto, M. M., Barnes, H. J., & Borst, L. B. (2018). A review of *Enterococcus cecorum* infection in poultry. *Avian Diseases, 62*(3), 261–271. https://doi.org/10.1637/11825-030618-Review.1

Justé, A., Lievens, B., Rediers, H., & Willems, K. A. (2014). The genus *Tetragenococcus*. In W. H. Holzapfel, & B. J. B. Wood (Eds.), *Lactic acid Bacteria: Biodiversity and taxonomy* (pp. 213–227). https://doi.org/10.1002/9781118655252.ch16

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ, 7, Article e7359*. https://doi.org/10.7717/peerj.7359

Knight, D. B., Rudin, S. D., Bonomo, R. A., & Rather, P. N. (2018). *Acinetobacter nosocomialis*: Defining the role of efflux pumps in resistance to antimicrobial therapy, surface motility, and biofilm formation. *Frontiers in Microbiology, 9*, Article 1902. https://doi.org/10.3389/fmicb.2018.01902

Kumar, S., Nei, M., Dudley, J., & Tamura, K. (2008). MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings in Bioinformatics, 9* (4), 299–306. https://doi.org/10.1093/bib/bbn017

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology, 5(2), Article R12*. https://doi.org/10.1186/gb-2004-5-2-r12

Langmead, B., & Salzberg, S. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods, 9*, 357–359. https://doi.org/10.1038/nmeth.1923

Le Crom, S., Schackwitz, W., Pennacchio, L., Magnuson, J. K., Culley, D. E., Collett, J. R., … Margeot, A. (2009). Tracking the roots of cellulase hyperproduction by the fungus *Trichoderma reesei* using massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America, 106*(38), 16151–16156. https://doi.org/10.1073/pnas.0905848106

Letunic, I., & Bork, P. (2024). Interactive tree of life (iTOL) v6: Recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Research, 52*(W1), W78–W82. https://doi.org/10.1093/nar/gkae268

Levy Karin, E., Mirdita, M., & Söding, J. (2020). MetaEuk-sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome, 8*(1), Article 48. https://doi.org/10.1186/s40168-020-00808-x

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics, 34* (18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics, 25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics, 25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lindstedt, K., Buczek, D., Pedersen, T., Hjerde, E., Raffelsberger, N., Suzuki, Y., … Sundsfjord, A. (2022). Detection of *Klebsiella pneumoniae* human gut carriage: A comparison of culture, qPCR, and whole metagenomic sequencing methods. *Gut Microbes, 14*(1), Article 2118500. https://doi.org/10.1080/19490976.2022.2118500

Liu, B., Zheng, D., Zhou, S., Chen, L., & Yang, J. (2022). VFDB 2022: A general classification scheme for bacterial virulence factors. *Nucleic Acids Research, 50*(D1), D912–D917. https://doi.org/10.1093/nar/gkab1107

Lu, J., Breitwieser, F. P., Thielen, P., & Salzberg, S. L. (2017). Bracken: Estimating species abundance in metagenomics data. PeerJ computer. *Science, 2, Article e104*. https://doi.org/10.7717/peerj-cs.104

Luo, Z., Zeng, W., Du, G., Chen, J., & Zhou, J. (2020). Enhancement of pyruvic acid production in *Candida glabrata* by engineering hypoxia-inducible factor 1. *Bioresource Technology, 295, Article 122248*. https://doi.org/10.1016/j.biortech.2019.122248

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution, 38*(10), 4647–4654. https://doi.org/10.1093/molbev/msab199

Matsen, F. A., Kodner, R. B., & Armbrust, E. V. (2010). Pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics, 11, Article 538*. https://doi.org/10.1186/1471-2105-11-538

Meyer, F., Fritz, A., Deng, Z. L., Koslicki, D., Lesker, T. R., Gurevich, A., … McHardy, A. C. (2022). Critical assessment of metagenome interpretation: The second round of challenges. *Nature Methods, 19*(4), 429–440. https://doi.org/10.1038/s41592-022-01431-4

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). MetaSPAdes: A new versatile metagenomic assembler. *Genome Research, 27*(5), 824–834. https://doi.org/10.1101/gr.213959.116

Oh, M., Han, J. W., Lee, C., Choi, G. J., & Kim, H. (2018). Nematicidal and plant growth-promoting activity of *Enterobacter asburiae* HK169: Genome analysis provides insight into its biological activities. *Journal of Microbiology and Biotechnology, 28*(6), 968–975. https://doi.org/10.4014/jmb.1801.01021

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

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research, 25*(7), 1043–1055. https://doi.org/10.1101/gr.186072.114

Price, M. N., Dehal, P. S., & Arkin, A. P. (2009). Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution, 26*(7), 1641–1650. https://doi.org/10.1093/molbev/msp077

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rambaut, A. (2016). *FigTree. version 1.4.3*. Institute of Evolutionary Biology, University of Edinburgh.

Raveendran, S., Parameswaran, B., Ummalyma, S. B., Abraham, A., Mathew, A. K., Madhavan, A., … Pandey, A. (2018). Applications of microbial enzymes in food industry. *Food Technology and Biotechnology, 56*(1), 16–30. https://doi.org/10.17113/ftb.56.01.18.5491

Salamzade, R., Cheong, J. Z. A., Sandstrom, S., Swaney, M. H., Stubbendieck, R. M., Starr, N. L., … Kalan, L. R. (2023). Evolutionary investigations of the biosynthetic diversity in the skin microbiome using IsaBGC. Microbial. *Genomics, 9(4), Article 000988*. https://doi.org/10.1099/MGEN.0.000988

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics, 24*(5), 637–644. https://doi.org/10.1093/bioinformatics/btn013

Stanton, R. A., Vlachos, N., & Halpin, A. L. (2022). GAMMA: A tool for the rapid identification, classification and annotation of translated gene matches from sequencing data. *Bioinformatics, 38*(2), 546–548. https://doi.org/10.1093/bioinformatics/btab607

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research, 27*, 737–746. https://doi.org/10.1101/gr.214270.116

Vendrell, D., Balcázar, J. L., Ruiz-Zarzuela, I., de Blas, I., Gironés, O., & Múzquiz, J. L. (2006). *Lactococcus garvieae* in fish: A review. *Comparative Immunology, Microbiology and Infectious Diseases, 29*(4), 177–198. https://doi.org/10.1016/j.cimid.2006.06.003

Von Wintersdorff, C. J. H., Penders, J., Van Niekerk, J. M., Mills, N. D., Majumder, S., Van Alphen, L. B., … Wolffs, P. F. G. (2016). Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Frontiers in Microbiology, 7, Article 00173*. https://doi.org/10.3389/fmicb.2016.00173

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., … Earl, A. M. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One, 9(11), Article e112963*. https://doi.org/10.1371/journal.pone.0112963

Warnow, T. (2013). SATe-enabled phylogenetic placement. In K. Nelson (Ed.), *Encyclopedia Of Metagenomics*. https://doi.org/10.1007/978-1-4614-6418-1_711-1

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

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with kraken 2. *Genome Biology, 20*(1), 257. https://doi.org/10.1186/s13059-019-1891-0

Xue, Y., Hu, M., Chen, S., Hu, A., Li, S., Han, H., Lu, G., Zeng, L., & Zhou, J. (2021). *Enterobacter asburiae* and *Pantoea ananatis* causing rice bacterial blight in China. *Plant Disease, 105*(8), 2078–2088. https://doi.org/10.1094/PDIS-10-20-2292-RE

Zheng, Z., Li, S., Su, J., Leung, A. W. S., Lam, T. W., & Luo, R. (2022). Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nature Computational Science, 2*(12), 797–803. https://doi.org/10.1038/s43588-022-00387-x