Development of strain-level shotgun metagenomics approaches to detect and characterize microbiological contaminants in the context of food safety

Florence Buytaers

Supervisors:
Prof. Dr. Ir. Kathleen Marchal (Ghent University)
Dr. Ir. Sigrid De Keersmaecker (Sciensano)

Members of the Examination Committee:
Prof. Dr. Sofie Goormachtig, chair (Ghent University)
Prof. Dr. Ir. Caroline de Tender, secretary (Ghent University)
Dr. Manal Abu Oun (Animal and Plant Health Agency (APHA), UK)
Prof. Dr. Marc Heyndrickx (Ghent University, ILVO)
Prof. Dr. Ir. Mieke Uyttendaele (Ghent University)
Prof. Dr. Kurt Houf (Ghent University)

Thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Science: Biochemistry and Biotechnology.

Academic year: 2022-2023

Development of strain-level shotgun metagenomics approaches to detect and characterize microbiological contaminants in the context of food safety

Florence Buytaers

Supervisors:
Prof. Dr. Ir. Kathleen Marchal (Ghent University)
Dr. Ir. Sigrid De Keersmaecker (Sciensano)

Members of the Examination Committee:
Prof. Dr. Sofie Goormachtig, chair (Ghent University)
Prof. Dr. Ir. Caroline de Tender, secretary (Ghent University)
Dr. Manal Abu Oun (Animal and Plant Health Agency (APHA), UK)
Prof. Dr. Marc Heyndrickx (Ghent University, ILVO)
Prof. Dr. Ir. Mieke Uyttendaele (Ghent University)
Prof. Dr. Kurt Houf (Ghent University)

Thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Science: Biochemistry and Biotechnology.

Academic year: 2022-2023

Ontwikkeling van shotgun metagenomics-methodes op stamniveau voor het opsporen en karakteriseren van microbiologische contaminanten in het kader van de voedselveiligheid

Florence Buytaers

Supervisors:
Prof. Dr. Ir. Kathleen Marchal (Ghent University)
Dr. Ir. Sigrid De Keersmaecker (Sciensano)

Leden van de examencommissie:
Prof. Dr. Sofie Goormachtig, chair (Ghent University)
Prof. Dr. Ir. Caroline de Tender, secretary (Ghent University)
Dr. Manal Abu Oun (Animal and Plant Health Agency (APHA), UK)
Prof. Dr. Marc Heyndrickx (Ghent University, ILVO)
Prof. Dr. Ir. Mieke Uyttendaele (Ghent University)
Prof. Dr. Kurt Houf (Ghent University)

Ingediend in gedeeltelijke vervulling van de vereisten voor de graad van Doctor in de Wetenschappen: Biochemie en biotechnologie.

Academiejaar 2022-2023

Summary

Foodborne contaminations are a global burden on public health worldwide. For many years, and until recently, the microbial contaminants have always been analyzed and characterized after isolation. Since the advent of second generation sequencing, it has been possible to characterize the complete genome of these isolates to the SNP level. It was proven that whole genome sequencing has a higher resolution than any combination of tests carried out so far. But the isolation is not always possible, and it is time-consuming. A new method based on the sequencing of all genetic material of the sample without isolation has become available some years ago, e.g. shotgun metagenomics. It allows to get a screenshot of every microbiological contaminant present in the sample at once, possibly also at the SNP level.

At the time this study started, shotgun metagenomics for the study of food contaminants was in its infancy. Moreover, strain-level characterization had only been achieved in a handful of studies, and had not been proven possible when more than one strain of the same species were present. Moreover, the genome obtained from metagenomics sequences had only rarely been associated with human cases of an outbreak, let alone during a real outbreak. Therefore, this PhD centered on the development of shotgun metagenomics methods to study microbiological foodborne contaminants to the strain level and with achievement of a relatedness study (phylogeny), with a focus on the applicability of the method in order to be easily implemented in reference laboratories in the future.

The method was first tested on minced beef spiked with shiga toxin-producing *E.coli* (STEC) at very low level (5 CFU/25g). After enrichment for 16 or 24 hours in buffered peptone water, the DNA was extracted with two classical commercial kits or with a kit performing depletion of the host DNA. The extracted DNA also was amplified or not using Phi29 polymerase. We showed that all sample preparation methods allowed to obtain a full characterization to strain level of the spiked strain in the beef sample, carrying another non-pathogenic strain of *E.coli*.

The simplest protocol was chosen for further studies (i.e. 24 hours enrichment as stated in the current international standard procedure, classical commercial DNA extraction and no amplification). Cheese samples were spiked with two different STEC strains, and we could cluster separately the reads corresponding to each strain and perform relatedness (phylogeny), however not all genes harbored in the isolate's genome could be retrieved when two strains of the same species were present.

The same protocol was then followed to investigate a real *Salmonella* foodborne outbreak. Two food samples were investigated and the Salmonella Enteritidis strain linked to the outbreak could be obtained, fully characterized, and related to the food and human isolates from the outbreak in a phylogenetic tree containing other Belgian sporadic cases and another *Salmonella* outbreak happening in Europe at the same time. Therefore, we could resolve the outbreak to its food source, and show the time saved (i.e. about two weeks) with shotgun metagenomics compared to the conventional methods.

The meat previously spiked with STEC was also used to investigate the difference between Illumina short reads or Oxford Nanopore Technologies (ONT) long reads sequencing. We showed that the same level of information could be obtained after sequencing with either of the two technologies, although ONT offered real-time sequencing, and 12 hours were enough to decipher the STEC strain from the endogenous *E.coli* strains while an Illumina MiSeq run takes 48 hours. Moreover, the lower-cost Flongle flow cell showed the same results after 24 hours of sequencing when using the host depletion DNA extraction method.

We then investigated the issue of the detection and characterization of genetically modified microorganisms (GMMs) in microbial fermentation products as a case study within the problematic of the spread of antimicrobial resistance in the environment. These organisms are genetically modified to enhance the production of a compound (e.g. enzymes, vitamins), and therefore a selection marker is often used to detect the bacteria who have included the modification in their genome. Antimicrobial resistance genes are often used as one of these markers. The construct may also include dependency to certain growing conditions, which hinders culturing or obtaining an isolate, in particular when the GMM is unknown. For these reasons, shotgun metagenomics was considered a good alternative to detect and characterize the contaminant, and no enrichment was conducted on these samples, that are considered as non-complex matrices as most of the DNA contamination should belong to the producing GMM if present. We showed that we were able to detect unnatural associations including AMR genes after sequencing all DNA in the samples, confirming the presence of a GMM and characterizing it.

Finally, we also investigated the contamination of food by viral pathogens such as norovirus and hepatitis A. In order to detect these RNA viruses, we extracted all RNA from the food (raspberry, bivalve). These samples were not enriched as it is particularly arduous to culture these viruses in laboratory conditions. Because the contamination level was low in a complex matrix, we tested several sample preparation methods that could enhance the detection of the virus in the sample or during the sequencing (i.e. adaptive sampling). Overall, we showed that shotgun metagenomics, with or without amplification, gave satisfactory results for moderate contamination levels (higher than 10^7 genome copies). Moreover, depending on the RNA extraction method, it might be used even for lower contamination levels (10^3 genome copies). Finally, a targeting of the norovirus by hybridization capture enhanced the relative quantity of reads classified as norovirus but at the expense of using a less open approach that can only characterize one viral species in the sample.

Overall, this thesis advanced the scientific knowledge about shotgun metagenomics for the study of food contaminants by attaining for the first time strain level resolution in samples with more than one strain of the same species, with both long and short reads sequencing technologies. Moreover, it offered proof of concepts of the feasibility of such a method, as asked by EFSA in a recent scientific opinion. This work also gave a precedent in outbreak resolution to the food source using metagenomics to the strain level, and detecting GMMs in microbial fermentation products. And finally, it presented clearly to the scientific community which sample preparation methods can or cannot allow to detect viral pathogens at low contamination levels in food samples without enrichment. All protocols that have been proposed have been chosen to be as close as possible to the ones currently used in the reference laboratories so they can be adapted to be used in routine in the future. Ultimately, a validation of the method is still necessary in order to obtain a precise limit of detection based on the analysis of a large dataset of samples. Moreover, other technologies can still be investigated to improve the results in particular when skipping the culture enrichment of the food and other applications for public health can be explored as well such as the analysis of the human microbiomes or the emergence of new contaminants.

Samenvatting

Voedselinfecties vormen een wereldwijde last voor de volksgezondheid. Jarenlang, en tot voor kort, werden de microbiële contaminanten altijd geanalyseerd en gekarakteriseerd na isolatie. Sinds de komst van de tweede generatie sequencing is het mogelijk het volledige genoom van deze isolaten te karakteriseren tot op SNP-niveau. Gebleken is dat sequencing van het volledige genoom een hogere resolutie heeft dan elke combinatie van tot nu toe uitgevoerde tests. Maar de isolatie is niet altijd mogelijk, en het is tijdrovend. Een nieuwe methode, gebaseerd op de sequencing van al het genetisch materiaal van het staal zonder isolatie, werd enkele jaren geleden beschikbaar, i.e. shotgun metagenomics. Hiermee kan in één keer een beeld worden verkregen van elke microbiële contaminant in het staal, eventueel ook op SNP-niveau.

Toen deze studie begon, stond shotgun metagenomics voor de studie van voedselcontaminaties nog in de kinderschoenen. Bovendien was karakterisering op stamniveau nog maar in een handvol studies bereikt, en was het nog niet mogelijk gebleken wanneer meer dan één stam van dezelfde soort aanwezig was. Bovendien was het uit metagenomics-sequenties verkregen genoom slechts zelden in verband gebracht met menselijke gevallen van een uitbraak, laat staan tijdens een echte uitbraak. Daarom richtte dit doctoraat zich op de ontwikkeling van shotgun metagenomics methoden om microbiologische voedselcontaminanten te bestuderen tot op stamniveau en met hierbij het kunnen uitvoeren van een verwantschapsstudie (fylogenie), met een focus op de toepasbaarheid van de methode om in de toekomst gemakkelijk in referentielaboratoria te kunnen worden geïmplementeerd.

De methode werd eerst getest op rundvlees gehakt waarin shiga toxine-producerende *E.coli* (STEC) op een zeer laag niveau (5 CFU/25g) artificieel was geïntroduceerd (‘spike’). Na aanrijking gedurende 16 of 24 uur in gebufferd pepton water werd het DNA geëxtraheerd met twee klassieke commerciële kits of met een kit die het DNA van de gastheer verwijdert. Het geëxtraheerde DNA werd ook al dan niet geamplificeerd met Phi29-polymerase. Wij toonden aan dat alle staalvoorbereidingsmethoden een volledige karakterisering tot op stamniveau mogelijk maakten van de gespikete stam in het rundvleesstaal, dat een andere niet-pathogene stam van *E.coli* bevatte.

Voor verdere studies werd het eenvoudigste protocol geselecteerd (d.w.z. 24 uur aanrijking zoals vermeld in de huidige internationale standaardprocedure, klassieke commerciële DNA-extractie en geen amplificatie). Aan kaasstalen werden twee verschillende STEC-stammen artificeel toegevoegd. We konden de reads van elke stam afzonderlijk clusteren en verwantschap vaststellen (fylogenie), maar niet alle genen in het genoom van het isolaat konden worden gevonden wanneer twee stammen van dezelfde soort aanwezig waren.

Hetzelfde protocol werd vervolgens gevolgd om een echte uitbraak van *Salmonella* in levensmiddelen te onderzoeken. Twee voedselstalen werden onderzocht en de *Salmonella*

Enteritidis-stam gerelateerd met de uitbraak kon verkregen en volledig gekarakteriseerd worden. Deze kon bovendien gerelateerd worden aan de voedsel- en humane isolaten van die uitbraak in een fylogenetische boom met daarin andere Belgische sporadische gevallen en een andere *Salmonella*-uitbraak die op hetzelfde moment in Europa plaatsvond. Hiermee konden we de uitbraak terugbrengen tot de voedselbron, en de tijdwinst aantonen (ongeveer twee weken) met shotgun metagenomics in vergelijking met de conventionele methoden.

Het eerder met STEC gespikete vlees werd ook gebruikt om het verschil te onderzoeken tussen Illumina short reads of Oxford Nanopore Technologies (ONT) long reads sequencing. Wij toonden aan dat hetzelfde niveau van informatie kon worden verkregen na sequencing met beide technologieën, hoewel ONT real-time sequencing aanbood, en 12 uur voldoende was om de STEC-stam te onderscheiden van de endogene *E.coli*-stammen, terwijl een Illumina MiSeq run 48 uur duurt. Bovendien toonde de goedkopere Flongle flowcel dezelfde resultaten na 24 uur sequencing bij gebruik van de gastheer-depletie-DNA-extractiemethode.

Vervolgens onderzochten wij de kwestie van de opsporing en karakterisering van genetisch gemodificeerde micro-organismen (GGM's) in microbiële fermentatieproducten als een casestudy binnen de problematiek van de verspreiding van antimicrobiële resistentie in het milieu. Deze organismen zijn genetisch gemodificeerd om de productie van een substantie (bijv. enzymen, vitaminen) te verhogen. Hiervoor wordt vaak een selectiemerker gebruikt om de bacteriën op te sporen die de modificatie in hun genoom hebben opgenomen. Antimicrobiële resistentiegenen worden vaak als een van deze markers gebruikt. Het construct kan ook afhankelijkheid van bepaalde groeiomstandigheden inhouden, wat het kweken of verkrijgen van een isolaat bemoeilijkt, met name wanneer het GGM onbekend is. Om deze redenen werd shotgun metagenomics als een goed alternatief beschouwd om de contaminant op te sporen en te karakteriseren. Er werd geen aanrijking uitgevoerd op deze stalen, die als niet-complexe matrices worden beschouwd, aangezien het grootste deel van de DNA-verontreiniging, indien aanwezig, tot het producerende GGM zou moeten behoren. Wij toonden aan dat wij onnatuurlijke associaties inclusief AMR genen konden opsporen na sequentiebepaling van al het DNA in de stalen, waardoor de aanwezigheid van een GGM werd bevestigd en gekarakteriseerd.

Ten slotte onderzochten we ook de besmetting van voedsel met virale pathogenen zoals het norovirus en hepatitis A. Om deze RNA-virussen op te sporen, extraheerden we al het RNA uit het voedsel (framboos, weekdieren). Deze stalen werden niet aangerijkt omdat het bijzonder moeilijk is om deze virussen in laboratoriumomstandigheden te kweken. Omdat het besmettingsniveau laag was in een complexe matrix, testten wij verschillende staalvoorbereidingsmethoden die de detectie van het virus in het staal of tijdens het sequencen (adaptive sampling) konden verbeteren. In het algemeen toonden wij aan dat shotgun metagenomics, met of zonder amplificatie, bevredigende resultaten gaf bij matige besmettingsniveaus (hoger dan 10^7 genoomkopieën). Bovendien kan het, afhankelijk van de RNA-extractiemethode, zelfs voor lagere besmettingsniveaus (10^3 genoomkopieën) worden gebruikt. Ten slotte verbeterde een gerichtheid op het norovirus door hybridisatie 'capture' de relatieve hoeveelheid sequencing reads die als norovirus werden geclassificeerd, maar dit ten koste van een open benadering, aangezien slechts één virale soort in het staal kan gekarakteriseerd worden met deze ‘capture’.

In het algemeen heeft dit proefschrift de wetenschappelijke kennis over shotgun metagenomics voor de studie van voedselcontaminanten bevorderd door voor het eerst een resolutie op stamniveau te bereiken in stalen met meer dan één stam van dezelfde soort, met zowel long-reads als met short-read sequencing technologieën. Bovendien leverde het een bewijs van de haalbaarheid van een dergelijke methode, zoals gevraagd door de EFSA in een recent wetenschappelijk advies. Dit werk gaf ook een precedent in het oplossen van uitbraken tot de voedselbron met behulp van metagenomics tot op stamniveau, en het opsporen van GGM's in microbiële fermentatie producten. En ten slotte heeft het de wetenschappelijke gemeenschap duidelijk gemaakt met welke staalvoorbereidingsmethoden virale pathogenen bij lage besmettingsniveaus in voedselstalen zonder aanrijking al dan niet kunnen worden opgespoord. Alle voorgestelde protocols werden zodanig gekozen dat ze zo gelijkaardig mogelijk zijn aan de huidige in referentielaboratoria geïmplementeerde protocols. Hierdoor kunnen ze in de toekomst aangepast worden om in de routine gebruikt te worden. Uiteindelijk is nog een validatie van de methode nodig om een precieze detectielimiet te verkrijgen op basis van de analyse van een grote dataset van stalen. Bovendien kunnen andere technologieën nog worden onderzocht om de resultaten te verbeteren, met name wanneer de cultuur-gebaseerde aanrijking van het voedsel wordt overgeslagen, en kunnen ook andere toepassingen voor de volksgezondheid worden onderzocht, zoals de analyse van het menselijke microbioom of de opkomst van nieuwe contaminanten.

Table of contents

List of abbreviations

AMR	Antimicrobial resistance
ARG	Antimicrobial resistance gene
BHI	Brain-Heart Infusion
bp	Base pair
BPW	Buffered Peptone Water
CFSAN	Center for Food Safety and Applied Nutrition
CFU	Colony Forming Units
cgMLST	Core-genome Multi-Locus Sequence Typing
cDNA	Complementary DNA
DALY	Disability adjusted life years
ddNTP	Dideoxyribonucleotides
DNA	DeoxyriboNucleotide Acid
dNTP	Deoxyribonucleotides
ECDC	European Centre for Disease Prevention and Control
EFSA	European Food Safety Authority
EU	European Union
EU-RL	European reference laboratory
FASFC (AFSCA/FAVV)	Federal Agency for the Safety of the Food Chain
FDA	Food and Drug Administration
FAO	Food and Agriculture Organisation
Gb	Gigabytes
GMM	Genetically Modified Microorganism
HAV	Hepatitis A virus
ISO	International Organization for Standardization
MAG	Metagenome-assembled genome
MALDI-TOF	Matrix-Assisted Laser Desorption/Ionization Time-Of-Flight
MLST	Multi-Locus Sequence Typing
MLVA	Multiple Locus Variable-number tandem repeat Analysis
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
NRC	National Reference Centre
NRL	National Reference Laboratory
ONT	Oxford Nanopore Technologies
PCR	Polymerase Chain Reaction
PFGE	Pulsed-Field Gel Electrophoresis
PHE	Public Health England
qPCR	Quantitative Polymerase Chain Reaction
RNA	RiboNucleic Acid

rRNA	Ribosomal RNA
SISTR	*Salmonella In Silico* Typing Resource
SNP	Single-Nucleotide Polymorphism
SRA	Sequence Read Archive
STEC	Shiga-toxin producing *Escherichia coli*
USA	United States of America
wgMLST	Whole-genome Multi-Locus Sequence Typing
WGS	Whole Genome Sequencing
WHO	World Health Organization

CHAPTER 1
Introduction

1.1. General context

1.1.1. The global burden of foodborne contaminants

Globally, every one person out of ten will get sick after eating contaminated food each year. Indeed, the World Health Organization (WHO) estimated that foodborne hazards cause approximately 600 million illnesses and 420,000 deaths per year worldwide (WHO, 2015; Lee and Yoon, 2021). Foodborne contaminations are a worldwide burden, not only for the health risk they represent, but also because of the economic impact caused by productivity loss. WHO created a parameter to estimate this overall burden, the Disability Adjusted Life Years (DALYs). This measure takes into account the years of life lost and the number of years to live with a disability, in this case due to the infection. The 31 foodborne hazards, including bacterial and viral foodborne pathogens, taken into account in their study summed up as 33 million DALYs, just for the year 2010. *Salmonella* and toxigenic *E. coli* were amongst the bacterial contaminants causing the highest burden while norovirus was the virus that caused the most illnesses. Interestingly, the same study was conducted in the USA in 2011, and concluded that 1 person out of 6 would get a foodborne infection each year (Scallan et al., 2011). The risk posed by foodborne hazards is therefore not restrained to people living in low-income regions.

Bacterial infections can theoretically be treated with antibiotics. However, overuse in humans, and preventive use in animals and the food-producing environment, combined with the large scale production of antibiotics leading to manufacturing waste and residues in the environment at sub-lethal dose, generated an increase in antimicrobial resistance (AMR), including in bacteria that can be found in food (EFSA-ECDC, 2022). The global issue of AMR represents a high impact that also can be calculated in DALYs: in 2019, it was estimated that bacterial AMR can be associated with 4.95 million deaths and almost 200 million DALYs yearly worldwide (Cassini et al., 2019; Murray et al., 2022). The potential for a bacterium to acquire resistance is greater when an antibiotic is present at low dose or sub-lethal concentration (Ching et al., 2020). The transfer of the antimicrobial resistance gene (ARG) can be vertical (i.e. from parent to offspring) or horizontal (i.e. transfer between different bacteria via conjugation, transformation, transduction or membrane vesicles). Notably, the horizontal transfer can occur even between distantly related species as it has been shown for plasmids, a circular DNA molecule found in microorganisms (Klümper et al., 2015). As a consequence, some multi-drug resistant strains have emerged, including in pathogens, and adversely influence the morbidity and mortality rate of patients undergoing some medical procedures (Prestinaci et al., 2015). Therefore, antimicrobial resistance is now recognized as a major global threat (Prestinaci et al., 2015; EFSA, 2021b).

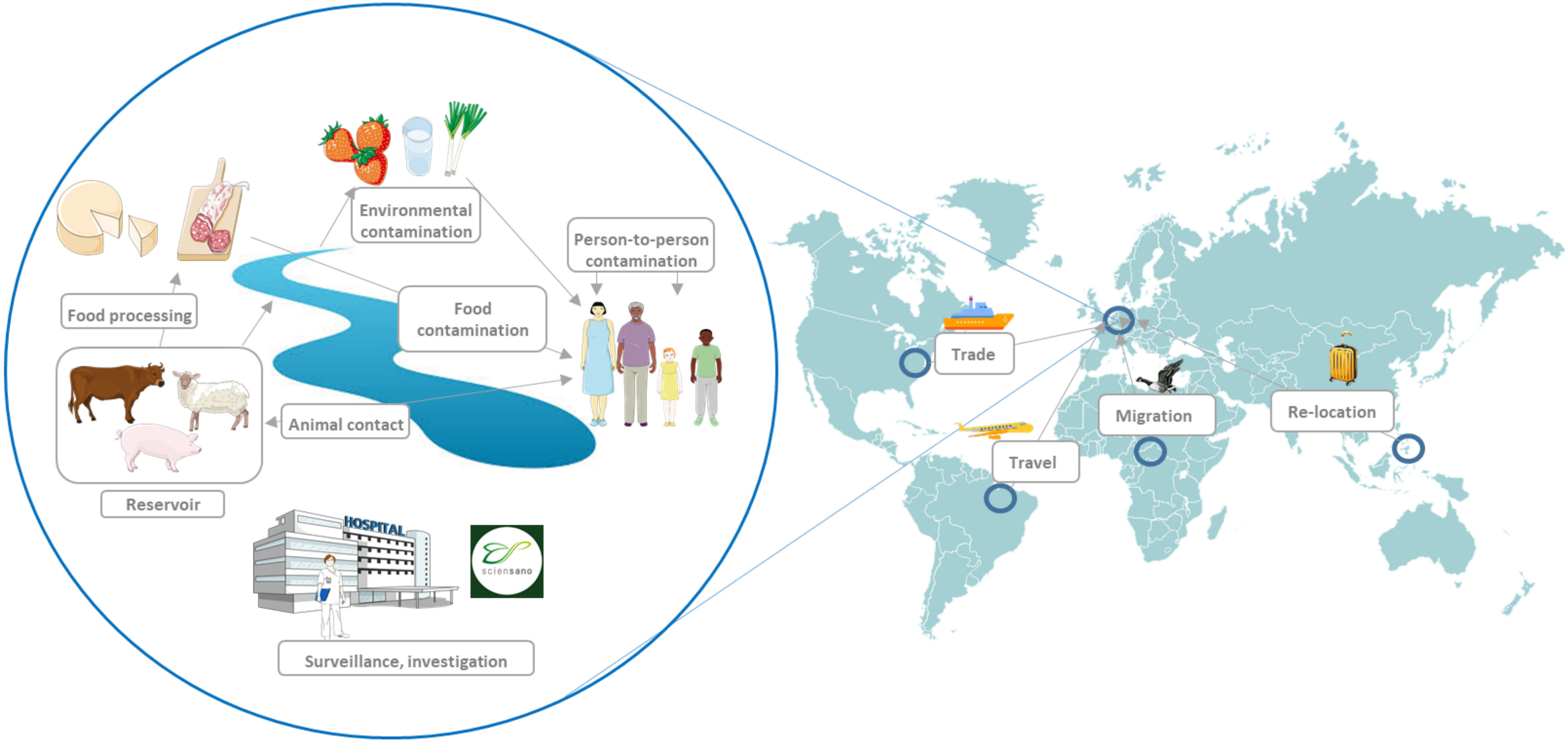


Figure 1.1: Potential foodborne contamination routes at a local (One Health) and at a global level. *Figure adapted from Hernando-Amado et al. and WHO (WHO, 2015; Hernando-Amado et al., 2019)*

1.1.2. The One Health approach in foodborne contamination control and surveillance

The food contaminants, including microorganisms containing ARGs, travel via different routes (Figure 1.1) before reaching the human population through consumption of food. This can be by consumption of a direct product of an animal (dairy, eggs, meat), but also by direc contact with an infected animal, for example at a petting farm. However, the animal product might first be processed at a manufacturing site. The contaminant can persist in a processing site and spread through food products transformed there. The human infections associated with an animal reservoir (often asymptomatic) are called zoonotic (EFSA, 2021c). People can also be infected through the environment (water, soil…) that can be contaminated by an animal carrying the pathogen (faeces…). This includes drinking contaminated water but also the bioaccumulation for example in shellfish (Nagarajan et al., 2022) or the contamination of fresh products such as leafy greens by irrigation water or water used for spraying insecticides (Okoh et al., 2010; Bottichio et al., 2020). Finally, infection can also occur by contact with a person that was previously infected (human-to-human transmission). This comprises also the contamination of food by contaminated food handlers (Bidawid et al., 2000; Somura et al., 2019). Some foodborne contaminants are preferably spread through one of these contamination routes, and they are not all zoonotic.

It is clear that there are several transmission routes possible, involving human, animal and environmental reservoirs. The One Health approach is a relatively new interdisciplinary concept based on the interdependence of human health, animal health and the environment, including the study of the food, through these transmission routes (EFSA, 2021c). Global Health (Hernando-Amado et al., 2019; Sinclair, 2019) is the connection of all One Health aspects at a worldwide level, influenced by travel, animal migration or trade for example (Figure 1.1).

Therefore, foodborne contamination control requires a One Health approach. Indeed, as shown with the transmission routes (Figure 1.1), foodborne pathogens can infect humans through animals, food products, the environment or contact with another infected person. Microorganisms containing ARGs also travel through the same routes. Several instances are in charge of the surveillance of these reservoirs. In Europe, the national reference centers (NCR) for the human samples, and the national reference laboratories (NRL) for the food and environmental samples, are the intersection between the local, regional and federal laboratories and authorities (Figure 1.2). They receive information and samples at a local and regional level from the medical sector, the clinical laboratories, the regional health authorities and the Regional Association for Animal Health and Identification (ARSIA/DGZ) and at a national level from the Federal Agency for the Safety of the Food Chain (in Belgium, AFSCA/FAVV), They also communicate at an international level to the European reference laboratories (EU-RL), the European Food Safety Authority (EFSA), the European Centre for Disease prevention and Control (ECDC) and the World Health organization (WHO).

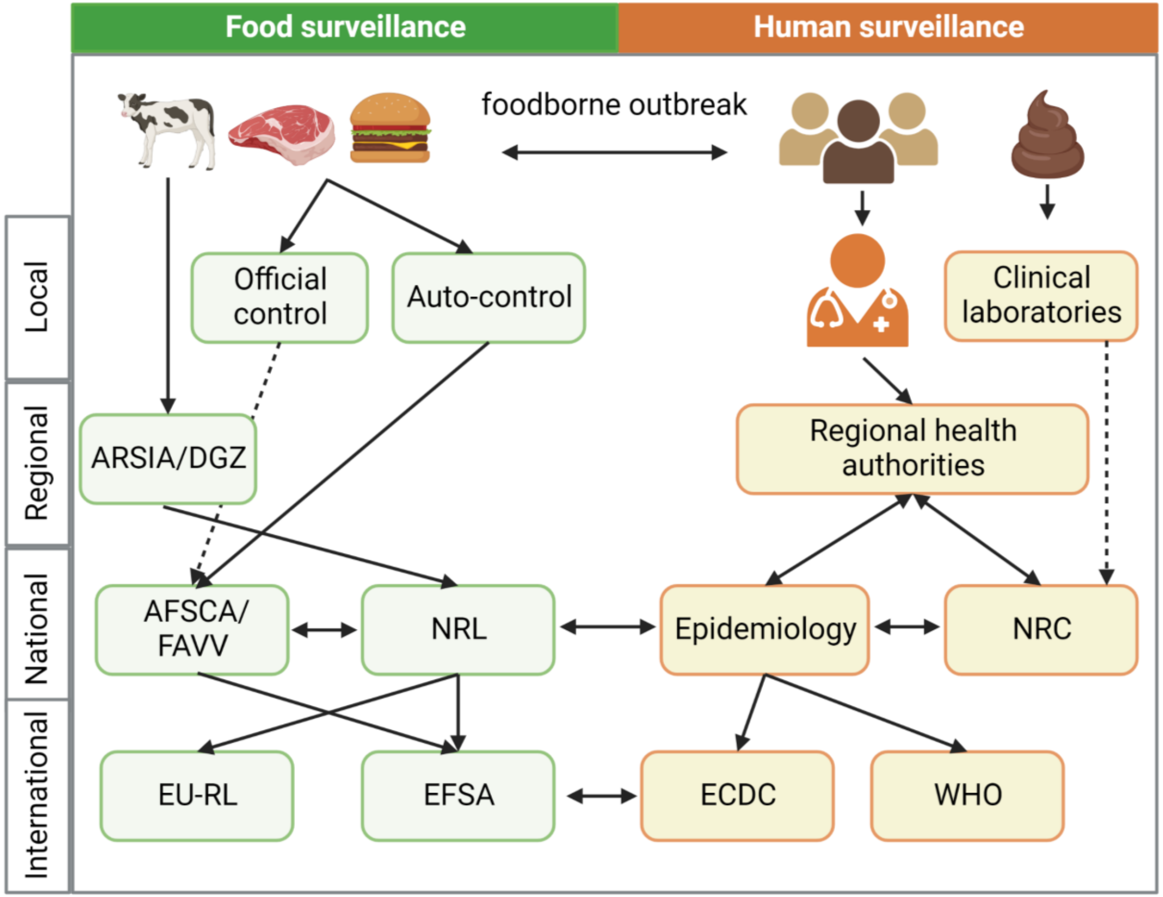


Figure 1.2: Overview of the information exchange during foodborne outbreak investigations in Belgium and in Europe. *Green: food surveillance. Orange: human surveillance. Arrow: data/sample exchange. Dotted arrow: voluntary data/sample sharing. Figure inspired by De Rauw, Gand, Sciensano and Uelze et al.* (De Rauw et al., 2019a; Gand et al., 2020; Sciensano, 2020; Uelze et al., 2021). *Figure made with Biorender.com*

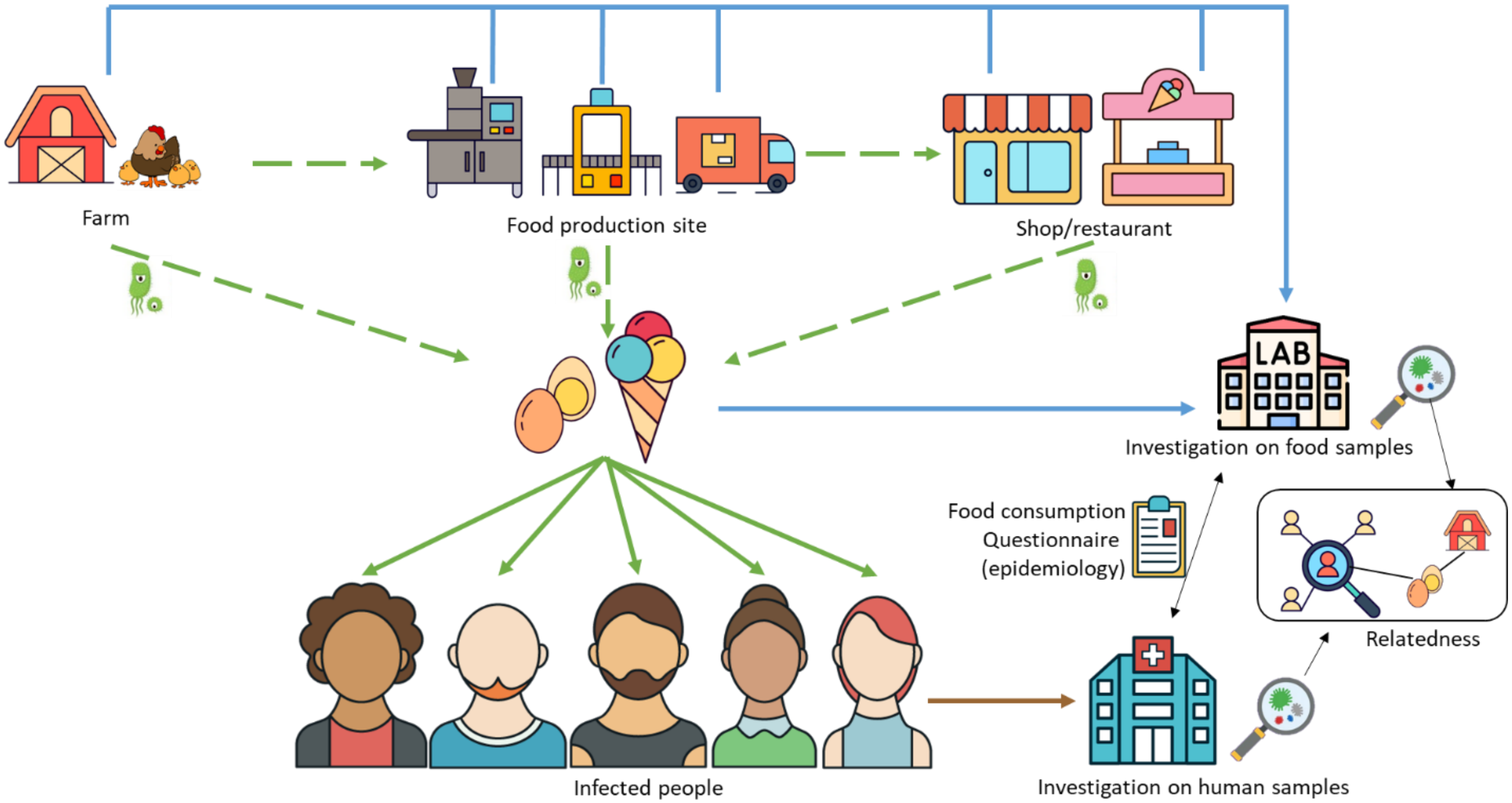


Figure 1.3: Overview of a foodborne outbreak from farm to fork (example from an egg farm). *Green: contamination route. Brown: sampling and analysis of the human samples. Blue: Sampling and analysis of the food/environmental samples. Black: exchange of information on the investigation on human and food samples to establish relatedness*

This communication is particularly active in case of an outbreak. A foodborne outbreak is declared when the same pathogen is detected in several people in a short time period (Figure 1.3). The information is centralized at the NRC and an epidemiological survey is conducted (through a food consumption questionnaire) to determine the food that might have caused this contamination. The suspect food is then sampled (Figure 1.3), along with possible investigation at the production site or at the farm (coordinated by FAVV/AFSCA in Belgium, Figure 1.2). Both human and food samples are tested to find the pathogen that caused the outbreak and characterize it. This information is gathered at the NRL for the food samples and the NRC for the human samples (Figure 1.2, 1.3). The NRC and NRL finally compare their reports. Based on the characterization data of the pathogens detected in the food and human samples, they can establish if there is a relatedness between the strains. This means that they can determine if the contaminant that caused the outbreak in the humans is the same as the one found in the food or environment (Figure 1.3). This is called the resolution of the food source of an outbreak, or trace back study, and it can lead back to the restaurant or shop where the food was purchased, food production site where the food was transformed or even the farm where it was produced (Figure 1.3). In some cases, foodborne outbreaks can spread across multiple countries, potentially also leading to more infections (ECDC, 2016; Hill et al., 2017). NRC and NRLs have communication channels within Europe to exchange at an international level (Figure 1.2). Many tests are available to detect and characterize pathogens within the food and human samples, but these need to obtain a sufficient level of information in order to present a strong evidence that the strains are related. EFSA defined this strong microbiological evidence : “Strong microbiological evidence includes the identification of an indistinguishable causative agent in a human case and in a food, a food component, or its environment, which is unlikely to have been contaminated coincidentally or after the event, or the identification of a causative agent, such as a toxin or bio-active amine, in the food vehicle, in combination with clinical symptoms and an onset of illness in outbreak cases strongly indicative/pathognomonic to the causative agent” (EFSA, 2014a) When this level of confidence is not attained, it is described as a weak-evidence outbreak resolution. When the contaminant cannot be characterized at all, it is described as an unknown agent.

1.2. Diverse spectrum of foodborne contaminants

1.2.1. Foodborne disease agents

The estimations on the burden of foodborne diseases presented above were calculated based on different foodborne hazards: biological agents such as viruses, bacteria, protozoa, helminths or chemical agents (Scallan et al., 2011; WHO, 2015). In this thesis, we will focus on microbiological contaminants. In particular, we will work on bacterial and viral agents.

In the 2015 WHO report on the burden of foodborne diseases (Figure 1.4.A), norovirus represented the highest count of illnesses per year worldwide (over 124 millions) while *Salmonella* Typhi and non-typhoidal *Salmonella enterica* caused the highest counts of deaths (over 50 thousand per year each) and highest DALYs. *Campylobacter* spp. and pathogenic or toxigenic *Escherichia coli* complete the top 5 (Figure 1.4.A). In Europe, *Salmonella* and norovirus are the most frequent sources of foodborne outbreaks (Figure 1.4.B) while Campylobacterioses, salmonelloses and shiga toxin-producing *E.coli* (STEC) infections are the most notified confirmed human zoonoses, as reported by EFSA in non-COVID circumstances (EFSA, 2021c). These results were obtained from the reporting of the NRL of each pathogen from the different countries of EU. Notably, strong-evidence outbreak resolutions are scarce compared to the weak-evidence outbreak resolution, and outbreaks caused by an unknown agent are placed in the top 3 of causes of foodborne outbreaks (Figure 1.4.B).

In this work, several case studies of foodborne pathogens have been investigated to deliver proofs of concept, based on high importance disease agents, i.e. *Salmonella enterica,* norovirus and Shiga toxin-producing *Escherichia coli* (STEC).

1.2.1.1. Salmonella enterica

The primary microbial agent causing foodborne outbreaks in Europe and worldwide is *Salmonella* (Figure 1.4.A and B). However, the associated disease (salmonellosis) has a low fatality rate of 0.19% (EFSA, 2021d). Two distinct species are part of the *Salmonella* genus: *Salmonella enterica* and *Salmonella bongori*, both pathogenic for humans. While *S. bongori* is rare and associated with cold-blooded animals (Fookes et al., 2011), *S. enterica* is a more common source of food infections and is divided in subspecies: *arizonae, diarizonae, houtenae, salamae, indica, and enterica* (Le Minor, 1988). *S. enterica* subsp. *enterica* is the most prevalent subspecies reported in Europe, further subdivided in more than 2700 serovars, with the serovar Enteritidis representing approximatively 50% of all *Salmonella* infections (EFSA, 2021d). *Salmonella* Enteritidis are the cause of zoonotic gastrointestinal diseases. The infectious dose is relatively high with approximatively 10^6 bacteria necessary to cause illness (Forsythe, 2000). The *Salmonella* genome, approximatively 5 million base pairs long, is also often harbouring antimicrobial resistance genes which can influence the response to the treatment of the patients and be transmissible into the environment (Thung et al., 2018). Although *Salmonella* can contaminate a variety of food matrices (e.g. fruit and vegetables, fish and fishery products, milk, meat…. (EFSA, 2021d)), the main source attribution of salmonellosis are eggs (Pijnacker et al., 2019). *Salmonella* are relatively easy to culture and can grow in a large range of temperatures and acidity (7-8 to 40°C approximatively and pH 4.5 to 9 (Uyttendaele, 2020)).

1.2.1.2. Foodborne viruses (norovirus and hepatitis A virus)

The second most common cause of foodborne outbreaks in Europe in non-covid conditions is norovirus (Figure 1.4.B, EFSA, 2021b) and it was reported to represent the highest burden of foodborne contaminations worldwide (WHO, 2015). It causes non-severe gastroenteritis and rarely leads to death (EFSA, 2021d). It is a positive-sense RNA virus with a genome of 7 kb. It has been divided in several genogroups, of which 3 are infecting humans (GI, GII and GIV). It is not a zoonosis and the infection is linked to the fecal-oral route. It is often linked to human-to-human contaminations, but has also been reported in bioaccumulating shellfish such as oysters (Strubbia et al., 2020), and it has been involved in

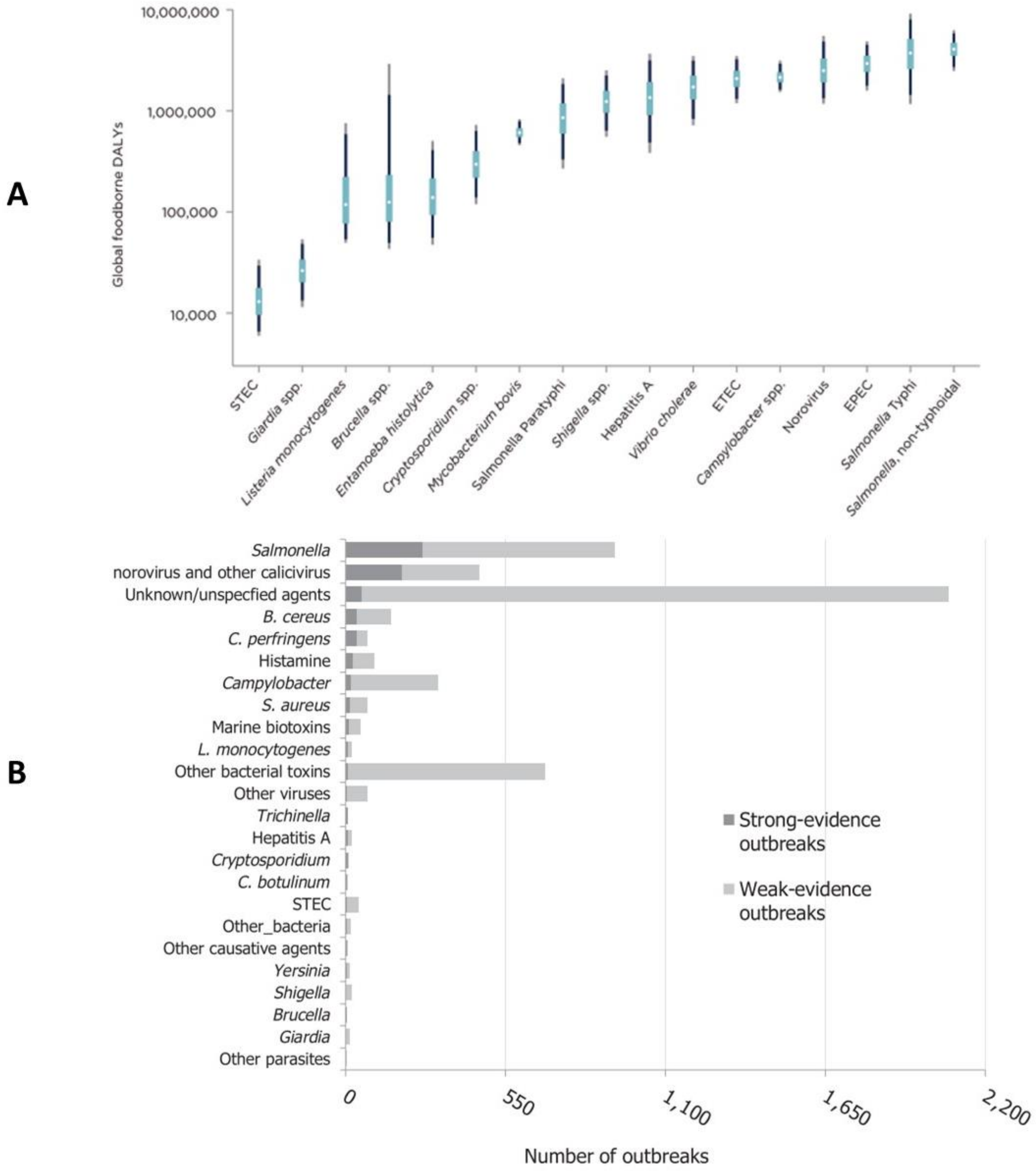


Figure 1.4: A: DALYs for each foodborne pathogen ranked from lowest to highest with 95% uncertainty intervals, 2010 (WHO, 2015). B: Distribution of outbreaks per causative agent in EU, 2019 (EFSA, 2021c)

outbreaks linked to the consumption of frozen berries that are handpicked (De Keuckelaere et al., 2015; Bartsch et al., 2018). The infectious dose is particularly low, with less than 100 viral particles potentially posing a risk (Yang et al., 2017) and it is very complex to culture viruses in laboratory conditions as it would require human cells (Jones et al., 2015).

Hepatitis A virus (HAV) is also a viral foodborne pathogen, but it is less common than norovirus and has a lower burden (WHO, 2015; EFSA, 2021c). It is a positive-strand RNA virus as well, with a genome of about 7 kb, and associated with contaminations through the fecal-oral route (Jeong and Lee, 2010). It has also been involved in large outbreaks linked to the consumption of shellfish (Halliday et al., 1991). Infection is asymptomatic in a large portion of the cases, especially in young children, but may lead to acute viral hepatitis. A vaccine is available since 1991 (Patravale et al., 2012), but is not mandatory.

1.2.1.3. Shiga toxin-producing Escherichia coli (STEC)

Escherichia coli are Gram-negative bacteria from the *Enterobacteriaceae* family, with a genome of about 5 million bases, that can be found along the gastrointestinal tract of humans and other animals. Most of them are commensal, but some can present pathogenic traits (Baker et al., 2016). It is the case for STEC, that acquired genes encoding the Shiga toxins (*stx1* and/or *stx2*) through integration of a phage (the stx phage). STEC is the fourth most frequent bacterial pathogen reported in Europe (EFSA, 2021d), and can lead to large outbreaks (Werber et al., 2012; Braeye et al., 2014) after consumption of only 10 to 100 bacteria (Feng et al., 2011). Eating undercooked meat or dairy products contaminated by the asymptomatic cattle carrying the pathogen, but also of other food products such as salads and sprouts sometimes contaminated through the water, can lead to diarrhea. But the infection can in some cases also end up in more serious conditions such as the hemolytic uremic syndrome (HUS). The presence of some genes other than the toxin encoding genes (e.g. *ehxA, eae, aaiC, aggR*) have shown a correlation to the seriousness of the disease (De Rauw et al., 2019b). STEC are divided in serogroups based on the antigen determination of an O and H type. The O157:H7 is the most common STEC associated with human infections (Butcher et al., 2016; Jajarmi et al., 2017), and the top 5 which represents most of the surveillance in Europe includes O157, O26, O103, O111 and O145 (EFSA, 2021d). STEC are able to grow and survive in a large panel of environmental conditions, however some strains and serotypes require more specific conditions to be cultured (Verhaegen et al., 2015).

1.2.2. Genetically modified microorganisms, a potential food contaminant harbouring ARGs

Food (or feed) products could not only be contaminated by pathogens but may also still contain for example microorganisms that were used for their production. This is the case for microbial fermentation products such as some food supplements like enzymes and vitamins. In that case, sometimes genetic modifications are introduced in the producing microorganism (in the genome or via introduction of a plasmid) in order to optimize the

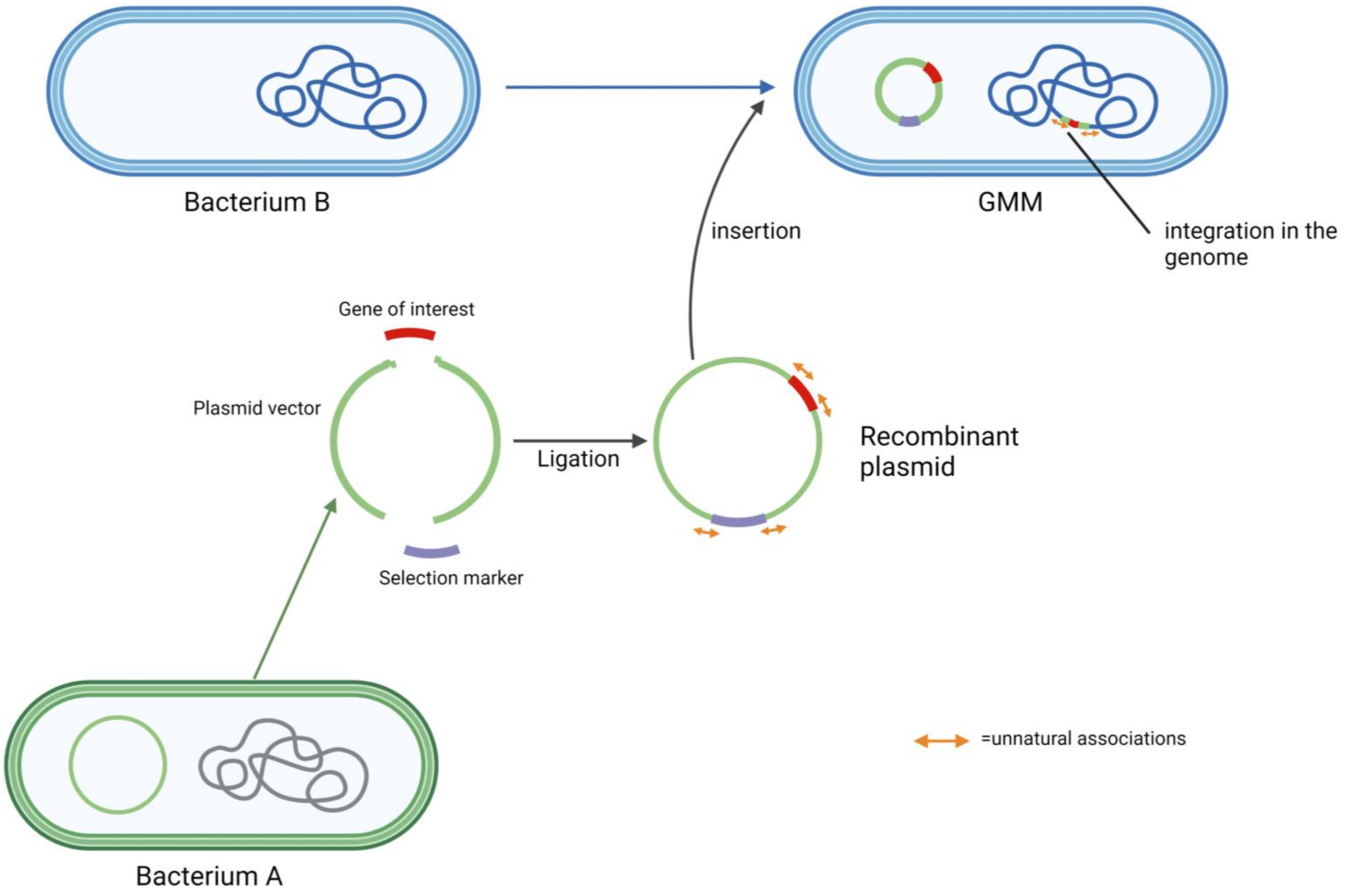


Figure 1.5: Schematic representation of the example of the construction of a genetically modified microorganism (GMM) based on a recombinant plasmid with a gene of interest and selection marker, incorporated as a plasmid in the producing bacterium or incorporated in the genome of the producing bacterium. *Figure made with Biorender.com*

manufacturing of the desired product (Deckers et al., 2020a). This is what we call a genetically modified microorganism (GMM, Figure 1.5). In order to detect the microorganism that assimilated the genetic modification during the modification process, selection markers are sometimes used. These selection markers are often the resistance to an antimicrobial agent. Therefore, all microorganisms that can grow on a media containing this antimicrobial have incorporated the genetic modification and will produce the desired enzyme. Moreover, some dependence to specific products in the culture environmnent can be introduced as well, leading to an auxotrophy. The introduction of genes from another bacterium in the genome of the modified microorganism induces the presence of unnatural associations (Figure 1.5). These unnatural associations can be found in the genome or on a plasmid introduced in the bacterium. Although according to EU regulations EC1331/2008, EC1332/2008 and EC1333/2008 (European Parliament and the Concil of the European Union, 2008; European Parliament and the Council of the European Union, 2008a, 2008b) the producing organism has to be removed from the final product, there have been reports of the accidental contamination of the food or feed product by the genetically modified microorganisms or its DNA (Barbau-Piednoir et al., 2016; Fraiture et al., 2020a)). These contaminants do not directly threaten the health of the consumers, but might represent a public health risk through the potential transfer of the acquired antimicrobial resistance to other species (WHO, 2018). Therefore, National Reference Laboratories are responsible for the screening of food and feed products for the potential presence of GMMs.

1.3. Evolution of methods for the detection, identification and characterization (typing) of microbiological contaminants in food samples

Several methods have been developed over the years to detect, identify and characterize the microbial contaminants in food samples. The various phenotypic tests and genotypic tests, which include PCR-based methods and sequence-based methods will be explained in this section.

1.3.1. The germ theory

Foodborne contaminants have had a close relationship with humanity since prehistory. Some have been found in Egyptian mummies (Hibbs et al., 2011) or microbiota of fossilized feces (Appelt et al., 2014) and it is even theorized that Alexander the Great died from water or food-borne contamination (Oldach et al., 1998). However, it was not before the 19th century that the link was made between the microorganism and the disease. In 1854, Dr. John Snow demonstrated that contaminated water was the source of the cholera epidemic (Snow, 1856). By the end of the 19th century, Dr. Louis Pasteur and other researchers confirmed the germ theory (Pasteur et al., 1879). They associated bacteria to diseases, and scientists started to try to find the source of a contamination at a microbiological level. This led to the discovery of many pathogens in the following decades, and the implementation of methods preventing the spreading of the diseases such as hygiene and pasteurization.

1.3.2. Phenotypic tests

The confirmation of the germ theory allowed to study microbiological foodborne contaminants, including those related to outbreaks, at the microbial level, by discovering the causative agent in the human samples and in the contaminating source. Another achievement from Dr. Louis Pasteur was the development of the first artificial liquid medium, allowing to culture microorganisms in a reproducible way. This medium was further optimized, leading to the development of a plethora of liquid or solid selective or non-selective media now available (Bonnet et al., 2020). These media enabled scientists to observe and count the microorganisms present in a sample, or a subset of them sharing some growing characteristics. These culture media allowed to discriminate an isolate, and therefore to identify or further characterize it. This was first done by observation of phenotypical traits. For example (non-exhaustive list), morphology examination under the microscope, Gram staining classifying bacteria based on their cell wall composition, biochemical tests such as the catalase test or later the analytical profile index (API), a miniaturized set of tests for which the compiled result can be compared to references, allowing to quickly identify relevant bacteria.The antibody typing, discriminating between different antigens present on the proteins at the surface of the microorganisms, led to the determination of the serotypes (Towner and Cockayne, 1993). The antibiotic susceptibility test (Dubourg et al., 2018; Pradhan and Tamang, 2019) is based on culturing in the presence of an antibiotic to determine from which concentration of the product the organism can grow in solid or liquid medium. Bacterial identification can also be conducted using gas chromatography analysis of the fatty acid methyl esters that have been extracted from whole cultured cells (GC-FAME, Tang and Row, 2013; Santos et al., 2018). The distribution of fatty acid methyl esters is dependent from one organism to another. However, these tests are often slow (Table 1.1), analyze only one bacterium at a time, and the microorganism do not always grow on the culture media (in particular viruses, but also some difficult to culture microorganisms such as GMMs). Moreover, several tests are necessary to identify and characterize the microorganism and the interpretation of the results can be subjective (Table 1.1).

Another phenotypic test which has been implemented as a high-throughput method to obtain a rapid identification of the isolated pathogen is based on matrix-assisted laser desorption/ionization-time-of-flight mass spectrometry (MALDI-TOF MS). This test allows the detection of the pathogen at a low price, and to analyze many samples at once in a short time-frame. Although this method improved the accuracy of the bacterial identification, the characterization of the strain is not possible, and the method is only comparable for species for which a validated database has been built (De Bruyne et al., 2011; Veloo et al., 2018).

Characteristic	Phenotypic tests			Molecular techniques		
	Microscopic techniques	Immunological techniques	Other phenotypic tests*	based on PCR	not PCR-based such as PFGE	Whole genome sequencing
Necessity for isolation	Not always +/-	Yes -	Yes -	Not always +/-	Yes -	Yes -
Time to result	5-30 minutes ++	10 min -1h (after isolation) +/-	Various (after isolation) +/-	1-3 hours ++	3 days (after isolation) -	3-4 days (after isolation) -
Sensitivity / resolution	1 cell/visual field ± 10^4 CFU/ml +/-	10^4-10^5 CFU/ml +/-	Various +/-	1 genome/PCR reaction +	+/-	SNP level ++
Specificity	-	+/-	-	+	+ (depending on the database)	++ (depending on the database)
Simultaneous multiparameter testing, including relatedness	+/-	-	- -	+	-	++
Differentiation of dead/viable cells (without taking isolation into account)	+/-	+/-	+/-	-	-	-
Reproducibility	+	+	+/-	+	+	+
Data analysis	+	+	+	+	+/-	Requires bioinformatics knowledge -
Interpretation of results	Human bias possible -	+/-	Human bias possible, need for databases +/-	+	+/-	+/-
Labour intensity	+/-	+/-	- -	+	-	+
Cost of materials	-	+/-	+	+/-	-	-
Investment of equipment	-	+/-	+/-	-	-	- -

Table 1.1: Perceived characteristics of conventional and newer methods for the characterization of microorganisms. **Other phenotypic tests described in this introduction e.g. API tests, maldi-TOF MS, GC FAME. The symbols describe 'overall perceived as a negative/positive/neutral intrinsic characteristic/(dis)advantage for the user', and more into detail: +: positive characteristic; + +: very positive characteristic; +/-: average/neutral characteristic; -: negative characteristic; - -: very negative characteristic. Adapted from Jasson et al. (Jasson et al., 2010)*

1.3.3. Genotypic tests

1.3.3.1. Polymerase chain reaction

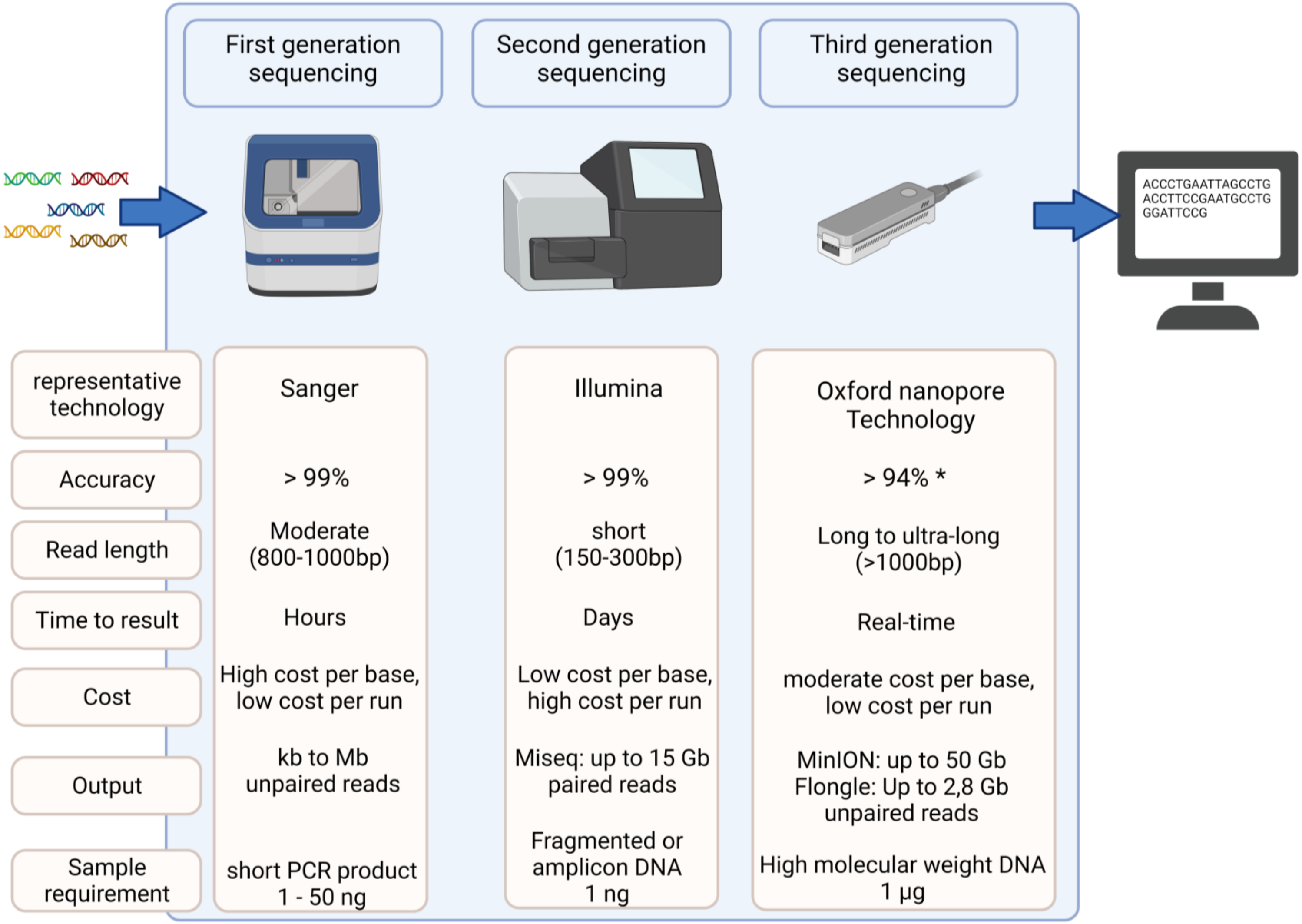
As exposed in the above section, the interpretation of the results of most of the phenotypic tests is subjective (Tang et al., 1998). Moreover, the accumulation of various tests to obtain a characterization of the isolate is fastidious (Table 1.1). A scientific breakthrough led to the development of fast and cost-effective genotypic tests (Kornberg et al., 1956; Kleppe et al., 1971), based on the characterization at the DNA level. With the polymerase chain reaction (PCR), the gene responsible for a specific phenotypic trait is detected after several cycles of amplifications. In addition, a real-time and quantitative monitoring of the DNA amplification, the real-time PCR (qPCR), was later developed based on the use of fluorescent dyes. These methods not only allowed to detect (and quantify) the DNA targets such as the strain type or virulence and antimicrobial resistance genes in isolates, but also directly in the matrices. Moreover, they offered a sensitive detection method (Table 1.1) in the case of viruses for which no culture was possible (Kralik and Ricchi, 2017). These methods became the gold standard in genetic testing for food microbiology and are still present in the protocols and international standards followed by the reference laboratories nowadays (ISO: International Organization for standardization, 2012, 2017).

1.3.3.2. Sequence-based methods

1.3.3.2.1. First generation sequencing

Although PCR methods allowed to unravel the presence of one or several genes within a genome, there was a need to develop methods to sequence the whole genome or whole genes in order to obtain the information at nucleotide level. The first generation of sequencing technologies (Figure 1.6) was made accessible when Frederick Sanger and colleagues developed the dideoxy chain-termination method (or Sanger sequencing), in 1977 (over twenty years after the discovery of the double helix structure by Watson & Crick). This technique was based on the use of technical analogues to the deoxyribonucleotides (dNTPs) lacking the 3’ hydroxyl group (dideoxyribonucelotides, ddNTPs), therefore preventing the formation of the 5’ phosphate bond to the next dNTP. Four parallel reactions of DNA extension were conducted with addition of one radiolabeled ddNTP in each of them (Sanger et al., 1977). The method was further improved by replacing radiolabeling by fluorescence detection, and using capillary based electrophoresis, to make the process automated and much faster. This allowed to produce the first sequencing machines.

Although this method has been exploited to sequence complete genomes (P Deloukas et al., 2001; Bentley et al., 2002), Sanger sequencing is reading one strand at a time and is mostly used to sequence smaller portions of genomes (hundreds base pairs, Figure 1.6). Therefore, it is often associated with PCR to sequence specific portions of a genome. It is used for determining the sequence of 16S rRNA gene which is commonly used for the identification of



	First generation sequencing	Second generation sequencing	Third generation sequencing
representative technology	Sanger	Illumina	Oxford nanopore Technology
Accuracy	> 99%	> 99%	> 94% *
Read length	Moderate (800-1000bp)	short (150-300bp)	Long to ultra-long (>1000bp)
Time to result	Hours	Days	Real-time
Cost	High cost per base, low cost per run	Low cost per base, high cost per run	moderate cost per base, low cost per run
Output	kb to Mb unpaired reads	Miseq: up to 15 Gb paired reads	MinION: up to 50 Gb Flongle: Up to 2,8 Gb unpaired reads
Sample requirement	short PCR product 1 - 50 ng	Fragmented or amplicon DNA 1 ng	High molecular weight DNA 1 µg

Figure 1.6: Different generations of sequencing technologies. **: accuracy depends on the reagents, flow cell and basecalling tool (Wang et al., 2021). Figure made with Biorender.com*

an organism in combination with reference databases (Loong et al., 2016) and has allowed a phylogenetic analysis revealing the repartition in three lines of descent, the eukaryotes, the bacteria and the archaea (Woese and Fox, 1977). It is also used e.g. for the typing of norovirus or in the context of GMMs to unravel unnatural associations in a genome, in combination with several PCR assays, in order to sequence the genomic regions flanking a known DNA sequence (i.e. based on DNA walking, Fraiture et al., 2015)

1.3.3.2.2. Second and third generation sequencing – high throughput sequencing

In the early 2000s, several new technologies such as 454 (later Roche), Sodexa (later Illumina) or Ion Torrent, each based on specific methodologies, were commercialized after two decades of research (Shendure et al., 2017). These new technologies, called the second generation (Figure 1.6) or next generation, supported massively parallel sequencing. They allowed to generate high amounts of sequences considerably faster than the process designed by Sanger (Mestan et al., 2011) as it would not sequence only one sequence at a time but many. High throughput sequencing allowed the characterization of the whole genome in one single test, providing a higher resolution in the characterization of microbes. This high throughput as well as the competition between the different instruments on the market lead to a rapid drop off in sequencing price per genome (or per bases), which is not yet over (Wetterstrand KA). Although various technologies have been developed based on different chemistries, Illumina sequencing has reached a near monopoly in the field (Heather and Chain, 2016). It is based on a sequencing by synthesis method after fragmentation. It relies on bridge amplification and reading of each nucleotide one by one using fluorophores (Figure 1.7). It allows to produce large datasets of paired short reads (100 to 300 bp depending on the chemistry) in runs of 48 to 72 hours with an accuracy averaging 99.9% (Sekse et al., 2017; Shendure et al., 2017, Figure 1.6).

In the next decade, a third generation of sequencing instruments emerged (Escobar-Zepeda et al., 2015). These methods are still considered under the umbrella of the “next generation sequencing” (Figure 1.6) but differ from their predecessors due to their ability to sequence single molecules, therefore generating longer reads. They also allow a sequencing and data analysis in real time (Eid et al., 2009). One of these approaches was developed by Pacific Biosciences (PacBio) and allowed since 2011 to observe via fluorescence the nucleotides integrated in the DNA chain during the polymerase reaction. This method allowed to obtain reads of several thousand base pairs and therefore opened new fields of possibilities, but it came at the cost of lower output and high error rate (about 10%) although randomly distributed (Shendure et al., 2017). It is also a big machine and a big investment like the second generation sequencing devices. In 2014, a new sequencing instrument was available for early access: The MinION from Oxford Nanopore Technologies (ONT). It is a very small device, in principle portable, and comes at a low cost. Its concept relied on the recording and translating into nucleotide sequence of the electric signal resulting from the flow of ions passing through the sequencing pore along with the DNA (or later developed also the RNA) fragment. This

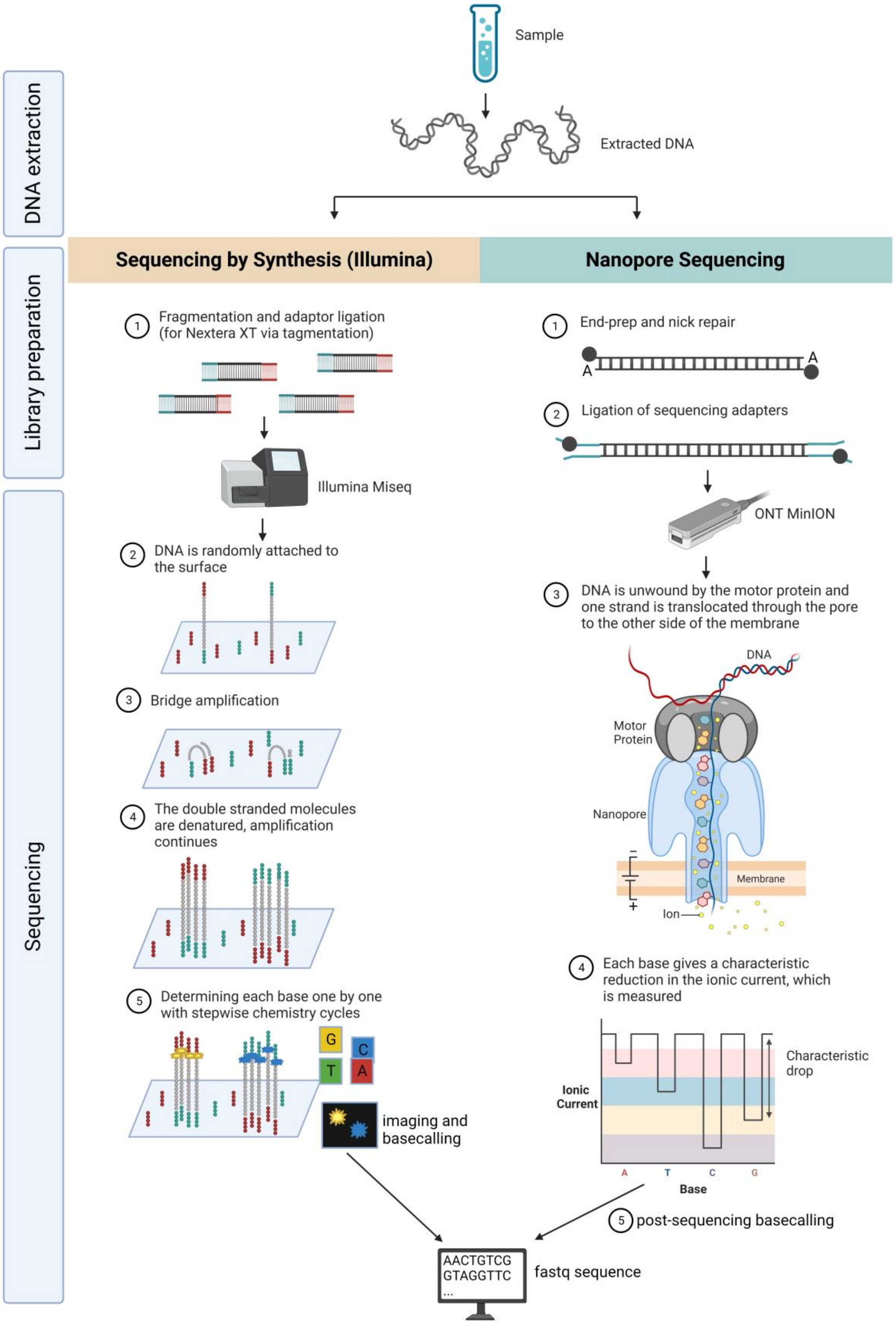

Sample
Extracted DNA
DNA extraction
Library preparation
Sequencing
Sequencing by Synthesis (Illumina)
Nanopore Sequencing
1 Fragmentation and adaptor ligation (for Nextera XT via tagmentation)
Illumina Miseq
2 DNA is randomly attached to the surface
3 Bridge amplification
4 The double stranded molecules are denatured, amplification continues
5 Determining each base one by one with stepwise chemistry cycles
G
C
T
A
imaging and basecalling
1 End-prep and nick repair
A
A
2 Ligation of sequencing adapters
ONT MinION
3 DNA is unwound by the motor protein and one strand is translocated through the pore to the other side of the membrane
DNA
Motor Protein
Nanopore
Membrane
Ion
−
+
4 Each base gives a characteristic reduction in the ionic current, which is measured
Ionic Current
Characteristic drop
A
T
C
G
Base
5 post-sequencing basecalling
AACTGTCG
GTAGGTTC
...
fastq sequence

Figure 1.7: Sequencing procedure for Illumina (Nextera XT tagmentation) and for Oxford Nanopore Technologies (ONT). *The DNA is extracted from the sample. Then the library is prepared following a protocol depending on the sequencing technology. For Illumina sequencing, a fragmentation is required, which can be conducted during the tagmentation if using the Nextera XT kit. The sequencing is then performed by using reversible dye-terminator sequencing-by-synthesis. Each individual DNA molecule is hybridized to the flow cell and the complementary DNA strand is created forming a double-stranded bridge. The fragments are amplified to form localized clusters on the flow cells from which the fluorescence corresponding to the nucleotide that has been added can be measured after each cycle. The basecalling is immediately done by the sequencing instrument and a fastq file is produced. For ONT's single molecule squencing, the library preparation does not necessarily include a fragmentation (optional). The information is obtained by translating the ionic current from the DNA fragment passing through the pore into nucleotide information. The basecalling is conducted post sequencing by the user to obtain a fastq file. Figure inspired by Zhou et al. and Loman et al.* (Zhou et al., 2010; Loman et al., 2012) *and made with Biorender.com.*

rapidly evolving technology showed a high error rate of around 10% at launch, but this has now decreased to 6% (Delahaye and Nicolas, 2021, Figure 1.6). Moreover, this instrument allows to sequence long and even ultra-long fragments of DNA, with the longest read recorded of 2.2 million bases and routine sequencing of fragments of several tens or hundreds of kbs (Payne et al., 2018). Lastly, another feature of the device is its portability, with the MinION being approximatively the size of a USB stick, that can be connected to a laptop to start a sequencing run.

Recently, a new type of flow cell, the Flongle, was launched by ONT as a low-cost alternative with about one tenth of the sequencing pores (between 50 and a hundred) and a consequently smaller output (Avershina et al., 2022, Figure 1.6). Other bigger devices are available (the GridION and the PromethION), corresponding to several MinIONs placed in parallel. A new feature has been presented on the GridION called “adaptive sampling” (denomination from ONT), which corresponds to a preferential sequencing of the reads corresponding to sequences in a database which is provided to the instrument prior to sequencing (Martin et al., 2022). ONT sequencing, however, requires 1000 times more DNA input and preferably high molecular weight (HMW) DNA to obtain longer reads, which can be challenging conditions.

Whatever the generation of the sequencer, whole genome sequencing (WGS) comes at a higher cost compared to the previously available methods (Table 1.1), but it allows to fully characterize the organism in just one test (Table 1.1). Generation of such big datasets initiated the new challenge to store and organize this data, and to obtain information that is easy to be interpreted (Table 8.1). Finally, as for all genotypic tests, WGS can only detect the presence of a gene, but it is not able to determine if the gene is expressed unlike the phenotypic tests.

Once the sequencing data is produced from the genetic material of the isolated contaminant, under the format of reads containing the nucleotide information, a data analysis is conducted to obtain the information on the organism that was sequenced (Figure 1.8). This can be done after first, second or third generation sequencing and on short or long reads, although the tools to use might vary depending on the sequencing technology. First, a quality check is conducted, which might include a profiling step to determine if the isolate’s DNA was contaminated with the DNA of another organism. The profiling or taxonomic classification consists in determining to which species belong the sequenced DNA or RNA and therefore if the isolate is of the expected organism, and if a contamination was present within the genetic material. This profiling can be conducted with various tools based on the use of different databases. They are usually referred to as DNA-to-DNA methods, DNA-to-protein methods and DNA-to-marker methods (Govender and Eyre, 2022). DNA-to-DNA methods such as Kraken2 (Wood et al., 2019) using for example the Refseq database (O’Leary et al., 2016) are widely used and score best when compared to the others (Ye et al., 2019; Govender and Eyre, 2022; Wright et al., 2022).

After the quality check, the reads can be filtered to retain only those with sufficiently good quality. Subsequently, the sequences can be analyzed as such or aligned to form longer fragments (contigs), and eventually reconstruct the genome of origin (Figure 1.8). Two

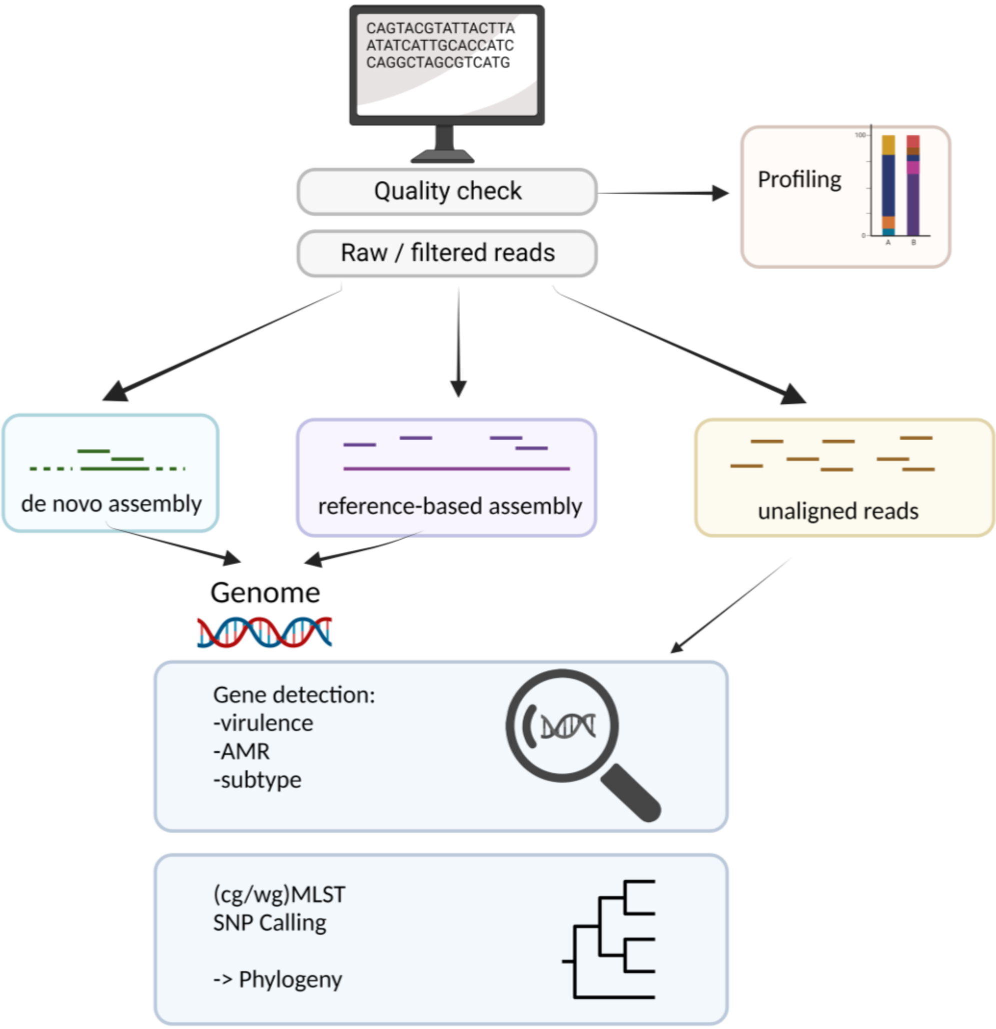


Figure 1.8: Various possible steps of a WGS data analysis workflow. *Figure made with Biorender.com*

alignment approaches are possible: the *de novo* assembly is based only on the overlapping of the reads without *a priori* information while the reference-based mapping consists of mapping the reads against a reference genome. The reads or contigs can then be characterized by mapping to a database of genes to detect if an element was present in the DNA and the degree of identity and coverage to the reference. In the case of the analysis of foodborne contaminants, the presence of ARGs within the genome of the contaminant can be determined by using databases such as ResFinder (Bortolaia et al., 2020). The pathogenic potential can be determined by detecting virulence genes using VirulenceFinder (Kleinheinz et al., 2014), while the serotype can be obtained based on serotyping genes (Joensen et al., 2015). All information from phenotypic and PCR-based tests can be obtained in one test.

1.3.3.2.3. Molecular-based methods to study relatedness between cases

No phenotypic method allows sufficient resolution to evaluate the relatedness between different strains of a same species that might or might not be linked. However, several molecular-based methods have been used for this purpose. The first developed method was the pulsed-field gel electrophoresis (PFGE (Arbeit et al., 1990)), a technique based on enzyme cutting sites used to separate large DNA molecules by periodically changing the direction of the electric field in a gel matrix. Although public databases are available with results of PFGE, this method has had limited reproducibility before the introduction of digital analysis and is relatively slow. It has been used as a gold standard for relatedness studies with the PulseNet database (Nadon et al., 2017). Multilocus enzyme electrophoresis or MLEE (Selander et al., 1986) is an alternative method based on the assessment of intracellular enzymes polymorphism using gel electrophoresis. PFGE has been shown to have better results for closely related isolates, for example in case of outbreaks, while MLEE is more appropriate to study further relationships and long term epidemiology of a contaminant (Tomayko and Murray, 1995; Maiden et al., 1998). Later, an approach based on PCR, the multiple loci variable number of tandem repeats analysis (MLVA (Keim et al., 2000)), has been proposed. It is based on the polymorphism in tandemly repeated sequences. It is a faster method and the profiles are easier to compare, but this method is not available for all pathogens. Moreover, it was shown that the comparison of PFGE or MLVA profiles can be misleading for the resolution of some foodborne outbreaks (Butcher et al., 2016; Nouws et al., 2021). Therefore, obtaining information at nucleotide level of the entire genome of the biological contaminant would give the highest resolution and discrimination as well as allowing to obtain all the informations from several tests at once. A first method based on the first generation sequencing of several housekeeping genes was presented, the multilocus sequence typing (MLST (Maiden et al., 1998)), and this was also used after whole genome sequencing with second or third-generation sequencers with an increased number of alleles to take into account (core genome (cg) or whole genome (wg) MLST). Finally, each nucleotide of the whole genome can be compared. Each nucleotide is compared between the genomes. A single nucleotide variant or single nucleotide polymorphism (SNP) is a variation of a single nucleotide that occurs at a specific position in the genome relative to the reference sequence. When SNPs are detected, if they pass specific filters to differentiate them from sequencing errors, this is referred to as SNP calling (Figure 1.8).

1.4. The current status of foodborne microbiological contaminants detection and characterization in NRLs and the need for a new method

1.4.1. Foodborne pathogens

Presently at the Belgian NRL for foodborne outbreak, when the pathogen is suspected to be present in the extracted DNA from a food sample after a qPCR test, the bacterial pathogen is attempted to be isolated from the food samples following enrichment using selective and non-selective media (Figure 1.9). If an isolate can be obtained, it is further typed using qPCR for specific markers (serogroup, virulence…) and/or phenotypic methods such as MALDI-TOF (ISO: International Organization for standardization, 2012, 2017). Depending on the contaminant, the trace back study can be performed using MLVA or PFGE (Figure 1.9).

In the case of viral pathogens, enrichment nor isolation is undergone, as it is not possible for most foodborne viruses. The virus is detected through qPCR in all reverse transcribed RNA extracted from the food sample. If the result is positive, a PCR is conducted for specific regions of the genome and these are Sanger sequenced to further type the contaminant (Figure 1.9, ISO: International Organization for standardization, 2019). No relatedness study is conducted.

Simultaneously with the technical developments of new characterization methods, recommendations have arisen from international agencies to use WGS (based on second generation sequencing) for major foodborne microbiological hazards (EFSA, 2013, 2014b; FAO, 2016; WHO, 2018). Indeed, obtaining the complete genome of a pathogen allows the highest level of characterization (Figure 1.9): the typing of the strain, the detection of virulence or antimicrobial resistance genes (also used for risk assessment), the detection or resolution of outbreaks, and high-resolution epidemiology (EFSA, 2014b; Tang et al., 2019). It also enables easy data sharing between institutions, even when different protocols are followed by the laboratories. Following these recommendations, sequencing-based typing of foodborne pathogenic isolates has gradually increased in Public Health Reference Laboratories for routine national surveillance (Revez et al., 2017; Van Goethem et al., 2020). It is now becoming the new standard method and its high level of sensitivity and specificity for outbreak investigations at local or international level has already been demonstrated in many studies (Butcher et al., 2016; Inns et al., 2017; Pijnacker et al., 2019; Nouws et al., 2020a). It has also been established that using WGS for routine surveillance is overall beneficial in terms of public health costs (Jain et al., 2019; Brown et al., 2021).

Although the increasing use of WGS as a characterization method allows to obtain the highest level of information, it is still not enough to solve the majority of foodborne outbreaks. In Europe we observe that the causative agent is identified in about 60 to 70% of the cases (EFSA, 2019a, 2021c, 2021d) but characterized (strong evidence outbreak resolution) only in

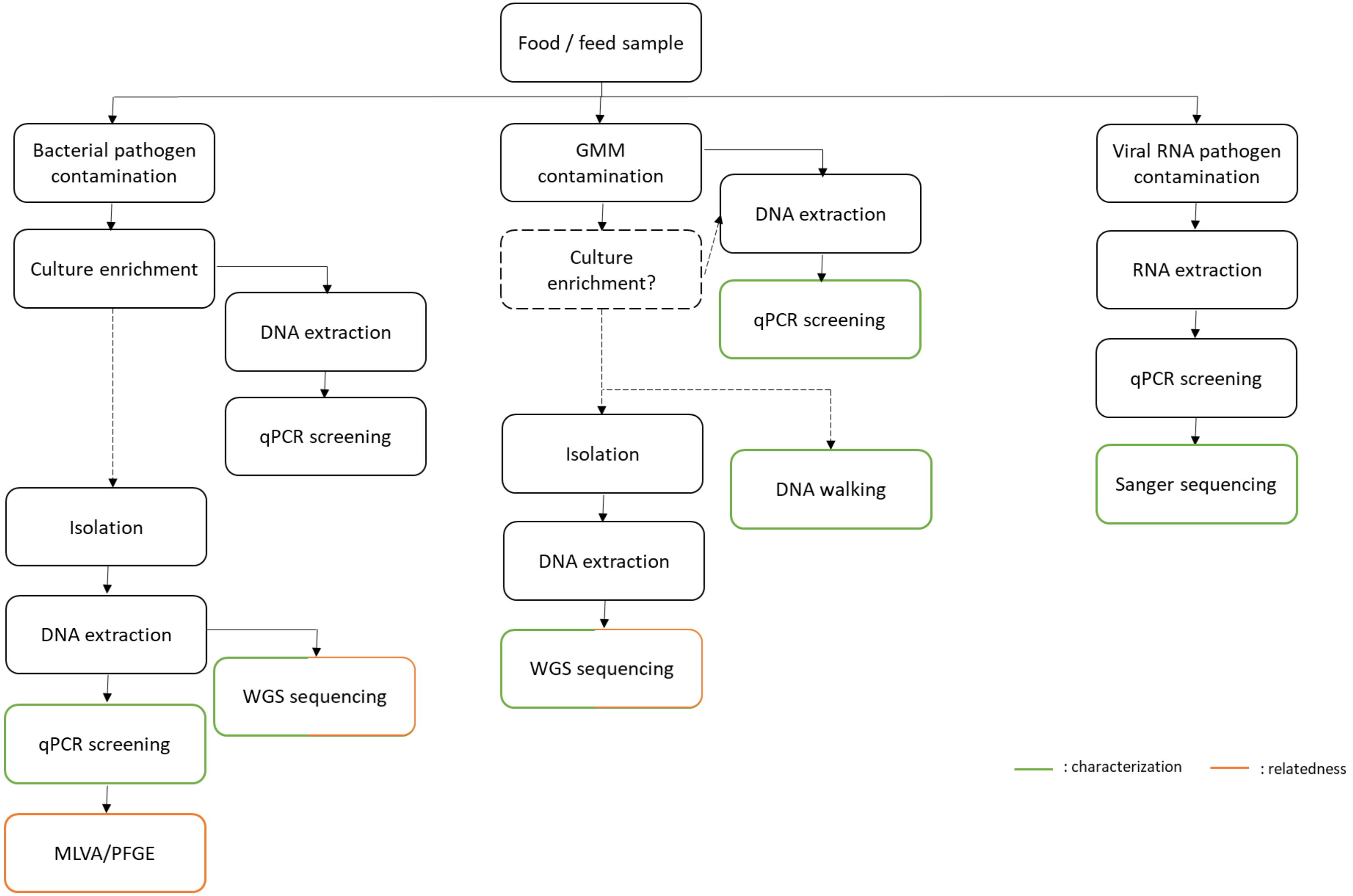


Figure 1.9: Overview of the current methods (conventional methods and WGS) used for the detection, characterization and relatedness of foodborne contaminants.

15% of these (EFSA, 2021c). For the remaining 85% of outbreaks, the resolution is based on weak evidence which means that the contaminant could not be isolated from the food in order to be characterized further, while it was however positively detected. In these high number of cases, no markers could be compared to confirm that it was the exact same strain that was present in the food and in the infected persons. Moreover, when the pathogen is not characterized, no risk assessment can be conducted on the pathogenicity or the transmission of antimicrobial resistances, but also the strain cannot be related to future or even historical cases and therefore the contamination cluster might not be fully cleared or followed up. Another important issue is the case of co-contaminations, by two strains of the same species or by two different species (Kinnula et al., 2018; Petronella et al., 2018; Liu et al., 2020), which might not always be correctly isolated to be sequenced later on. These have been shown to occur in some outbreaks, but the current isolation procedures might prevent from detecting another contaminant present in the sample in most cases.

1.4.2. Genetically modified microorganisms

The procedure used by the NRL for the identification and characterization of GMMs in food samples is based on the same methods and technologies (Figure 1.9), but with a different goal. Currently, the DNA of the sample is extracted with or without culture enrichment of the sample, and two rounds of qPCR screening are executed: a first-line screening for the detection of ARGs typically found in known GMM constructs (Fraiture et al., 2020b), a vector also associated with publicly accessible GMM patents (Fraiture et al., 2020e) or the *Bacillus subtilis* group, a species often used as GMM (Fraiture et al., 2022). It is followed by a second-line screening for previously characterized (event-specific) GM constructs (Barbau-piednoir et al., 2015; Fraiture et al., 2020a, 2021a, 2021b). Notably, at the time of writing, only three GM constructs (i.e. *B. subtilis* producing protease (Barbau-piednoir et al., 2015), *B. velezensis* producing protease (Fraiture et al., 2021a) and *B. amyloliquefaciens* producing alpha-amylase (Fraiture et al., 2021b)) are known and can be screened for. When a sample gives a positive signal for first-line screening but not for second-line screening, it can potentially be an unknown GMM. One way of characterizing the contaminant is to attempt to obtain an isolate. If the isolate can be cultured, a WGS analysis can be performed (Berbers et al., 2020). The WGS data can then be used to detect unnatural associations and then develop new second-line screening methods (Barbau-piednoir et al., 2015; Fraiture et al., 2020a). When the isolate cannot be obtained, another way to characterize the contaminant and assess if it is a GMM, is to conduct DNA walking around genes of interest (Fraiture et al., 2017, 2020e) such as the ARGs and shuttle vectors positive in the first-line screening. This is however only possible with *a priori* information, i.e. primers to anchor the sequencing, and it often requires several sequential tests to obtain sufficient information to confirm the presence of an unnatural association.

Accordingly, the current situation is not effective enough as only a handful of specific screening tests are currently available (the second-line qPCR tests), while the vast majority of recombinant organisms stay undetected, with their sequences unknown.

1.5. Metagenomics as an alternative to the conventional methods

1.5.1. The metagenomics approach

Because isolation is not always possible for all organisms and might not offer a correct overview of all the microbes present in a sample due to the enrichment conditions, microbiologists have for long wanted to characterize the whole genome of the various microorganisms directly in the sample at nucleic acid level (Escobar-Zepeda et al., 2015). This would allow to study microbial communities, discover novel organisms or to observe the species present in the food. The main bias is the uncertainty about the living state of the organism from which DNA was extracted (Quince et al., 2017).

A first approach to obtain information on the bacteria present in samples, including food samples, i.e. metabarcoding, was developed as PCR targeting of the 16S ribosomal RNA gene, used to identify microorganisms present in a sample (Grützke et al., 2019; Deckers et al., 2020c). This method was first developed to identify isolated organisms (see chapter 1.3.3.2.1) and was later used without isolation. It can only target organisms harbouring a 16S rRNA gene (bacteria, archea), while fungi and eukaryotes are characterized by the ITS or 18S rRNA regions, respectively. With this method only a small part of the genome is sequenced, which does not allow to obtain all the genomic information about the organisms present in the samples. Nonetheless, this approach is much less expensive approach than a whole genome sequencing, and is therefore used to profile the microorganisms present in a sample. The identification at species level, however, has been reported as challenging (Winand et al., 2019).

Later, another approach, i.e. shotgun metagenomics, or what will be referred to as "metagenomics" in the remainder of this work, was developed to sequence all the genetic material without any selection. This includes the matrix (human, animal or plant DNA depending on the sample), and all the microorganisms living on it. DNA extraction is conducted on the entire sample, and these nucleic acids are all sequenced. RNA extraction can also be conducted followed by direct RNA sequencing or sequencing of the complementary DNA (cDNA) after reverse transcription of the RNA, and is associated to the study of the transcriptome (metatranscriptomics) but also the analysis of the genomes of RNA viruses (Rajagopala et al., 2021).

When characterizing genomes, both methods rely on the same data analysis workflows. A profiling of the sequenced reads (determination of the species present in the sample) can be conducted in order to have a view on all the species originally present in the extract of the sample. It is done by comparing the metagenomics reads to a database of reference sequences, most often of whole genomes (Quince et al., 2017). A commonly used tool for taxonomic classification is Kraken2 (Wood et al., 2019; Govender and Eyre, 2022). When used with metagenomics datasets, it is often combined with a database of full genomes (such as Refseq (O'Leary et al., 2016)). This tool can also be used on isolates for the quality check as previously described.

As previously stated, the advantage of the shotgun metagenomics compared to the metabarcoding or qPCR is the access to the entire genetic information and the genetic context (presence of a gene on a specific genome). The goal is therefore to re-obtain the genetic information of each strain, based on the sequenced reads, and to be able to characterize these genomes the same way as if the genome had been sequenced from an isolate (Figure 1.8). In order to do this, the reads have to be re-attributed to each species or strains originally present in the sample. The reconstruction can be performed using databases of previously sequenced microorganisms (reference-based analysis) or without any information (*de novo* assembly). In the case of low contaminations (low coverage of the contaminant), reference-based classifiers are most appropriate. Tools like Sigma will classify each read to the most closely related reference from the database used (Ahn et al., 2015). All reads classified as one or several closely related reference genomes can be considered a strain and further characterized. This is however not appropriate when the contaminant is unknown and therefore not present in the databases. Then, a *de novo* assembly is more appropriate. The obtained strain/genome after reference-based or *de novo* analysis can then be characterized such as an isolate after WGS, e.g. with gene detection, in order to obtain for example, the virulence profile, the resistance to antibiotics or the subtype. Moreover, the strain/genome can be compared to other cases in a relatedness study. Therefore, it can help in resolving outbreak investigations by linking the strains from food and human origin and would offer a valid alternative to the characterization of the isolate with conventional methods or with WGS (Figure 1.9).

In order to detect sequences of the contaminant within the mix of reads, it is important to have this contaminant in the DNA extract. Therefore, several sample preparation methods can be used (Figure 1.10). Before DNA extraction, the food can be incubated (culture enrichment) in selective or non-selective media for various durations and at different temperatures, in order to enrich for possible biological contaminants. After DNA extraction (for which many kits or methods are available), the genetic material can be amplified in order to increase the quantity of DNA that can be sequenced, including the reads of the contaminants (Conceição-Neto et al., 2015). Finally, a selection can be operated at the level of the genetic material with the aim to increase the sequencing of reads corresponding to the contaminants. This can be done with various methods with various degrees of openness, targeting or depleting the genetic material that is not the contaminant (depletion of eukaryotic DNA (Grützke et al., 2021), depletion of rRNA (Liefting et al., 2021), targeting of poly-adenylated RNA (Fonager et al., 2017), capture of specific targets (Brown et al., 2016; Hyeon et al., 2018)). Finally, adaptive sampling is also an option to increase the number of sequenced reads corresponding to the pathogen or to decrease the sequenced reads corresponding to the genetic material of the matrix.

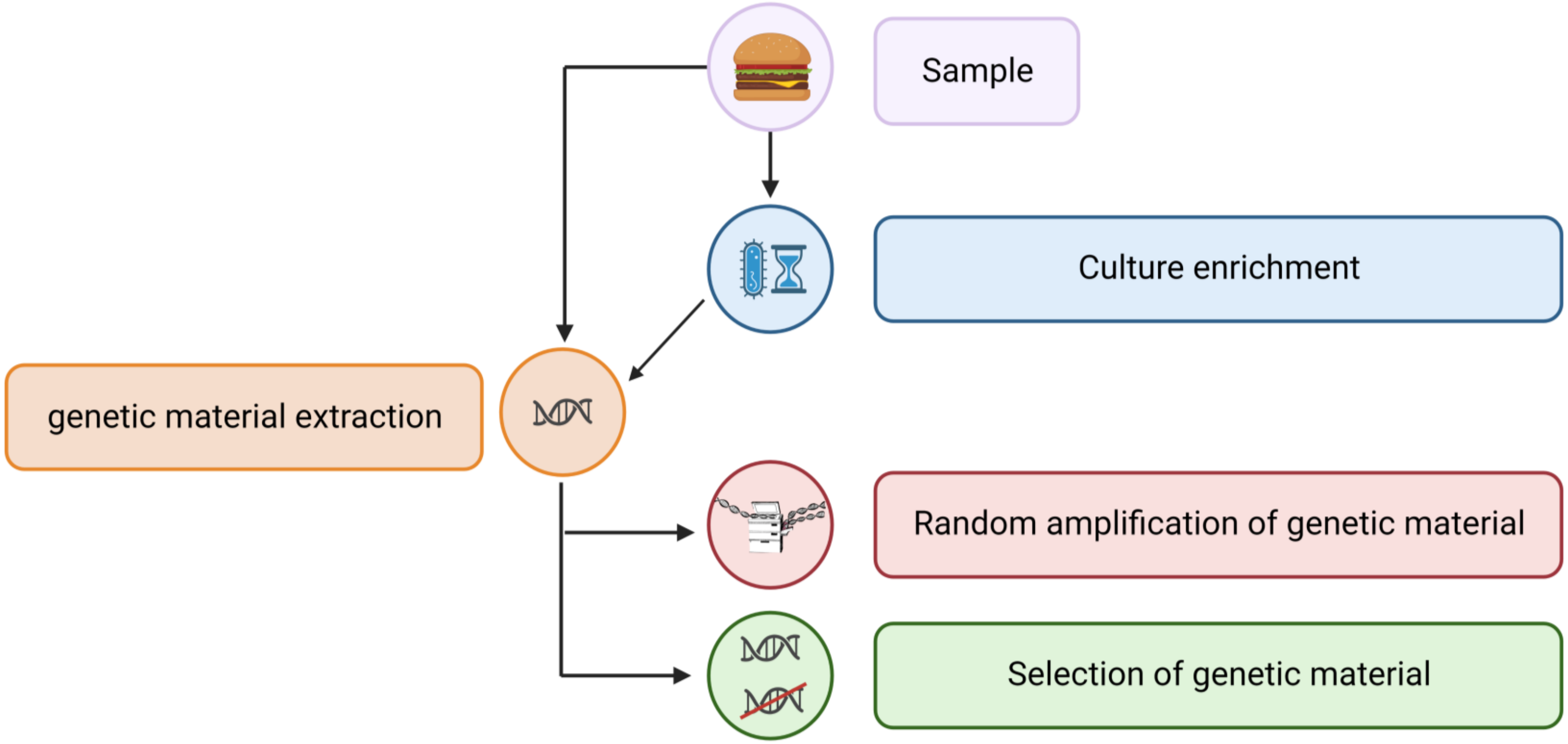

Sample
Culture enrichment
genetic material extraction
Random amplification of genetic material
Selection of genetic material

Figure 1.10: Various preparation methods used for metagenomics samples: the extraction of the genetic material can follow or not a culture enrichment step*. If the concentration is very low, the genetic material can be amplified e.g. by random DNA amplification. If the contamination is low, a selection of the genetic material can be conducted to increase the portion of the contaminant within the extract Figure made with Biorender.com*

1.5.2. Current situation of shotgun metagenomics for the study of foodborne contaminants

Metagenomics was first presented in 1996 (Stein et al., 1996) and the terminology was proposed by Handelsman in 1998 (Handelsman et al., 1998). The method was at that time mostly used for environmental studies to investigate the microbiome, including non-culturable organisms, in environments that were not yet fully described. Therefore, when this PhD started, shotgun metagenomics for the study of food contaminants was still in its early development while conventional isolation-based or 16S metabarcoding were more widely adopted.

The first level of information that was shown to be obtained from metagenomic sequencing in food samples was the identification of the species using reads of the entire genome instead of only using a small part of the genome (metabarcoding). In 2011, Park and colleagues used shotgun sequencing (Roche 454), producing a few tens of thousands reads to look at the viral communities in fermented food (Park et al., 2011). In 2012, Kawai et al. studied unresolved outbreaks linked to the consumption of raw fish, and were able after shotgun metagenomic sequencing to identify *Kudoa septempunctata* as the causative agent (Kawai et al., 2012). Later, teams applied shotgun metagenomics to identify potential human and animal viruses in fresh produce (lettuce, parsley leaves) and irrigation water, working with both RNA and DNA that were amplified with non-biased PCR (Aw et al., 2016; Fernandez-Cassi et al., 2017).

Some studies went further and also used the metagenomics reads to detect genes that were present in the mix. In 2012, Nieminen and his team looked at the bacteria linked to spoilage in different meats but also at the associated metabolic genes using Roche 454 sequencing (Nieminen et al., 2012). Yang et al. used Illumina sequencing to identify the microbial species along the beef production chain and looked at the potential presence of a pathogen by detecting virulence factors in the sequenced reads (Yang et al., 2016).

However, the genes detected in these studies were not linked to a genome. In 2014, Zhang and colleagues were able to detect and partially characterize viruses present in beef, pork and chicken meat after non-specific amplification and Illumina MiSeq sequencing (Zhang et al., 2014). They could detect protein-coding genes and open reading frames (ORFs) in the viruses. The following years, Leonard et al. reported in two consecutive studies the characterization at strain level of a bacterial pathogen (STEC) in spinach spiked at very low level (0.1 CFU/g) after 8 hours of enrichment with specific antibiotics, and Illumina sequencing (Leonard et al., 2015, 2016).

Lastly, the strain-level information with relatedness to other case studies was first described for foodborne pathogens after working with faeces of infected patients, such as the study of Loman and colleagues in 2013, who presented a STEC foodborne outbreak investigation based on metagenomics sequencing (Loman et al., 2013). They were able to assemble *de novo* the genome of the outbreak strain, map reads of each metagenomics sample to this genome, and look for the presence of particular genes such as the *stx* genes in each of these samples. Notably, in these fecal samples, the STEC outbreak strain accounted for a large proportion of the total sequenced reads, while a STEC contamination in a food sample, which was not investigated, would represent a few hundreds to a few thousand genome copies. Other teams worked on the detection and characterization of viruses in faeces and required special sample preparation methods increasing the viral load in the sample (Figure 1.10): Fonager et al. used poly(A) capture (Fonager et al., 2017), Van Beek and colleagues used hybridization with custom baits (Van Beek et al., 2017) and Nasheri et al. used non-specific amplification (Nasheri et al., 2017). Finally, this was also attempted on food samples with the work of Yang et al. and Bartsch et al. on foodborne viruses, with no specific sample preparation but with a targeted bioinformatics analysis after Illumina sequencing (Yang et al., 2017; Bartsch et al., 2018) while Walsh et al. analyzed the microbial content of a fermented beverage to the strain level and performed phylogeny on the detected strains with the same sequencing technology (Walsh et al., 2017). Hyeon et al. found relationships between their quasimetagenomics-based strains after long reads sequencing (Hyeon et al., 2018). However, they targeted the pathogen (*Salmonella*) with immunomagnetic separation followed by DNA amplification, which is not a totally open approach (therefore the name quasimetagenomics) and they still showed relatively high levels of SNPs between closely related cases.

For the case of the GMM contaminations, at the time this PhD started, very few studies had been published. Only one GMM had been previously isolated and fully characterized with WGS (Barbau-Piednoir et al., 2016; Paracchini et al., 2017). A qPCR screening was developed for this specific GM construct (Barbau-piednoir et al., 2015), and a DNA walking strategy based on PCR and Sanger sequencing was also available to possibly investigate other cases (Fraiture et al., 2015, 2017). The other qPCR methods used for the screening (Chapter 1.4.2 and Figure 1.9) were not yet available and shotgun metagenomics had never been used for this case study.

1.5.3. Challenges to implement shotgun metagenomics as an alternative method to study biological foodborne contaminants

Although at the time this PhD research started, the work already published was very promising, several challenges still had to be overcome to use shotgun metagenomics at the strain level for the characterization of foodborne contaminants and the resolution of outbreaks at the same level as the conventional methods. First of all, proofs of concept were necessary. These proofs of concept should show that strain-level metagenomics could be attained, even at low levels of contamination in food samples, with a fully open approach, and allow to determine relatedness to linked cases. The results should prove to be comparable to those of WGS on isolates, from food and human origin.

The approach should also be evaluated on several food matrices and contaminants. Moreover, depending on the contaminant, the level of contamination will vary due to the lowest infective dose but also the possibility to enrich the sample by culturing. When no culture enrichment is possible (GMMs, viruses), obtaining sufficient reads for a characterization of the genome of the contaminant is very challenging. Therefore, different sample preparation methods (Figure 1.10) might have to be investigated. The extraction of the genetic material should also be evaluated, in particular on different food matrices, as it might impact the species that are preferentially extracted and therefore the detection of the contaminant within the genetic material and sequenced reads. Another added value of metagenomics is its possible ability to characterize more than one strain in a sample with only one test. However, this is very arduous and had not yet been achieved when this PhD started.

The choice of the sequencing technology will possibly also have an impact on the level of information that can be obtained from the contaminant. Illumina was widely used when this work started, including for shotgun metagenomis studies. But Oxford Nanopore Technologies, as well as the low-cost Flongle flow cells, had not yet been fully compared for a strain-level metagenomics analysis. Notably, the use of another sequencing technology will impact the laboratory protocols and data analysis tools to be used, which should be carefully investigated.

Improving the time and cost to obtain the characterization of the contaminant would have an important impact on the field, and shotgun metagenomics might contribute to this progress. However, it has not yet been carefully evaluated. The real-time sequencing technologies (ONT) could offer a solution, although the error rate must be carefully taken into account to determine if strain-level characterization with low levels of contamination is achievable with this sequencing technology.

Finally, a new protocol will only be implemented in a routine environment if it is able to detect the contaminants at the low levels comparable to those detected with the current methods, stated in the current international standards and norms. This would facilitate future possible accreditation of the method. Furthermore, the developed methods should get out of the research setting to become more easily applicable in a routine laboratory. For this, the sample preparation and genetic material extraction should be as close as possible to the current methods in the reference laboratories, and follow normal working hours' conditions, while still obtaining a high level of characterization. Such practical protocols have not been proposed yet. Moreover, achieving to study a real foodborne outbreak, with relatedness to human cases, if possible, would convince the authorities and scientific community of the strength of this method in real conditions.

CHAPTER 2
Scientific problems addressed and aims of the thesis

2.1. Scientific problems addressed

Rapid and accurate characterization of the microbiological composition of a food sample allows the detection of the biological contaminants that might be present in it, and lead to the prompt management of a foodborne outbreak. In recent years, a reflection has begun on the use of next-generation sequencing (NGS) methods to improve outbreak investigations. Indeed, it allows to obtain the complete DNA sequence and to infer relevant information such as subtype, presence of virulence genes or antibiotic resistance factors, as well as tracing the origin of the contamination, at single nucleotide resolution. Therefore, NGS is set to become the new standard for rapid characterization of contaminants in food microbiology. The focus of this endeavour has, however, mostly been on the use of whole genome sequencing (WGS) of pathogenic isolates, whereas the process of obtaining an isolate from food samples is often time-consuming and not straightforward.

Metagenomics is an alternative approach based on the sequencing of all genetic material in the samples, that requires no isolation, and can be applied for the full genomic characterization of foodborne bacteria but also viruses or parasites. However, it is a very new approach and proofs of concepts still have to be established for its ability to characterize food contaminants at strain level.

The aim of this PhD research was to develop and deliver a proof of concept for a method able to characterize, if possible, to the strain level, a biological contaminant present in food samples without prior isolation, i.e. a shotgun metagenomics-based approach. In the scope of this work, several research questions have been investigated:

- Which metagenomics approach (sample preparation, sequencing, bioinformatics analysis) could allow characterization and relatedness at least at the same level as the conventional methods?
- How does the contamination load and/or matrix influence the approach to be followed?
- How to adapt the approach depending on the biological contaminant?
- How does the sequencing technology influence the results?
- How to achieve fast and cost-efficient results?
- How to implement this approach in routine analyses?

2.2. Outline of the thesis

The outline of the different chapters of this thesis is presented in Figure 2.1. The first chapter provided an introduction of the general context, the food contaminations, the current methods to detect and characterize them in order to provide safe food, the need for a new approach and a presentation of the status of shotgun metagenomics as an alternative approach at the time this doctoral research started, with several challenges to overcome. The second chapter states the aim and the outline of the thesis and several scientific questions to answer during this work. Chapter 3 describes the development of a strain-level metagenomics approach for short-reads sequencing applied to the case of a bacterial contaminant spiked at a low contamination level in an enriched food matrix. Chapter 4 presents the application of the workflow developed in Chapter 3 on a real foodborne outbreak. Chapter 5 states the adaptations to the workflow introduced in Chapter 3 when using a long reads sequencing approach and compares the outputs and cost-effectiveness of the two approaches on the same samples. Chapter 6 describes how shotgun metagenomics can be used to detect and partially characterize a GMM present in a non-complex food matrix, without enrichment, using short or long reads sequencing. Chapter 7 outlines different approaches to characterize foodborne viruses present at low contamination doses without enrichment in complex matrices using long reads sequencing. Finally, Chapter 8 gives a general discussion and some conclusions and perspectives on this thesis.

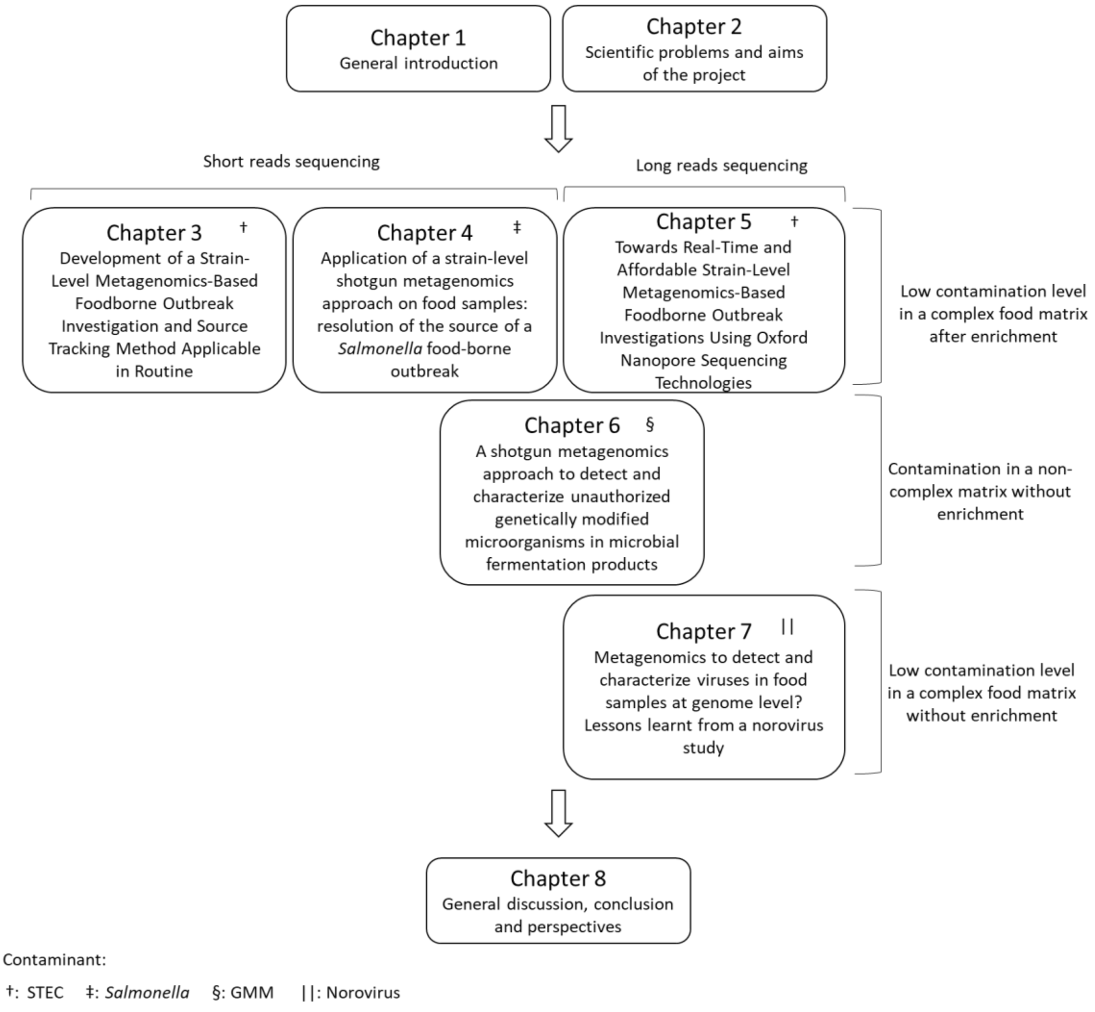

Chapter 1
General introduction
Chapter 2
Scientific problems and aims of the project
Short reads sequencing
Long reads sequencing
Chapter 3 †
Development of a Strain-Level Metagenomics-Based Foodborne Outbreak Investigation and Source Tracking Method Applicable in Routine
Chapter 4 ‡
Application of a strain-level shotgun metagenomics approach on food samples: resolution of the source of a *Salmonella* food-borne outbreak
Chapter 5 †
Towards Real-Time and Affordable Strain-Level Metagenomics-Based Foodborne Outbreak Investigations Using Oxford Nanopore Sequencing Technologies
Low contamination level in a complex food matrix after enrichment
Chapter 6 §
A shotgun metagenomics approach to detect and characterize unauthorized genetically modified microorganisms in microbial fermentation products
Contamination in a non-complex matrix without enrichment
Chapter 7 ||
Metagenomics to detect and characterize viruses in food samples at genome level? Lessons learnt from a norovirus study
Low contamination level in a complex food matrix without enrichment
Chapter 8
General discussion, conclusion and perspectives
Contaminant:
†: STEC ‡: *Salmonella* §: GMM ||: Norovirus

Figure 2.1: Outline of the thesis.

CHAPTER 3
A Practical Method to Implement Strain-Level Metagenomics-Based Foodborne Outbreak Investigation and Source Tracking in Routine based on STEC as a case study

This chapter was adapted from the previously published manuscript:
Buytaers, Florence E.*, Assia Saltykova*, Sarah Denayer, Bavo Verhaegen, Kevin Vanneste, Nancy H. C. Roosens, Denis Piérard, Kathleen Marchal, and Sigrid C. J. De Keersmaecker. 2020. "A Practical Method to Implement Strain-Level Metagenomics-Based Foodborne Outbreak Investigation and Source Tracking in Routine" Microorganisms 8, no. 8: 1191. https://doi.org/10.3390/microorganisms8081191. (* equal contribution)

<u>Authors’ contributions:</u>

F. E. Buytaers designed the study and methodology, produced and analysed the data, interpreted the results and drafted the manuscript. A. Saltykova helped to develop the data analysis workflow and interpret the results. S. Denayer, B. Verhaegen and D. Piérard curated bacterial isolate collection and provided bacterial isolates. K. Vanneste was responsible for providing the data analysis infrastructure. S. Denayer, B. Verhaegen, D. Piérard, K. Vanneste, N. H. C. Roosens and K. Marchal provided specialist feedback. S. C. J. De Keersmaecker conceived and supervised the study, helped to design the study, to interpret the results and to draft the manuscript. All authors have read and approved the manuscript.

Abstract:

The management of a foodborne outbreak depends on the rapid and accurate identification of the responsible food source. Conventional methods based on isolation of the pathogen from the food matrix and target-specific real-time polymerase chain reactions (qPCRs) are used in routine. In recent years, the use of whole genome sequencing (WGS) of bacterial isolates has proven its value to collect relevant information for strain characterization as well as tracing the origin of the contamination by linking the food isolate with the patient's isolate with high resolution. However, the isolation of a bacterial pathogen from food matrices is often time-consuming and not always successful. Therefore, we aimed to improve outbreak investigation by developing a method that can be implemented in reference laboratories to characterize the pathogen in the food vehicle without its prior isolation and link it back to human cases. We tested and validated a shotgun metagenomics approach by spiking food pathogens in specific food matrices using the Shiga toxin-producing *Escherichia coli* (STEC) as a case study. Different DNA extraction kits and enrichment procedures were investigated to obtain the most practical workflow. We demonstrated the feasibility of shotgun metagenomics to obtain the same information as in ISO/TS 13136:2012 and WGS of the isolate in parallel by inferring the genome of the contaminant and characterizing it in a shorter timeframe. This was achieved in food samples containing different *E. coli* strains, including a combination of different STEC strains. For the first time, we also managed to link individual strains from a food product to isolates from human cases, demonstrating the power of shotgun metagenomics for rapid outbreak investigation and source tracking.

3.1. Introduction

Food contaminations with pathogens are a major burden on our society, affecting an estimated 600 million people a year and impacting socioeconomic development at various levels (WHO, 2015). Microbial contaminations include bacteria, viruses, or parasites and regularly result in extensive outbreaks as foodstuffs can be processed and traded at a large scale. In case of foodborne outbreak investigation, the microbiological analysis of the probable responsible food vehicle is performed at two levels and consists of the detection of the pathogen, followed by the association of the food vehicle to the human cases using typing of the food isolate. The fast and accurate source attribution allows to remove the product from the market and limit its impact on the population. For the detection of bacterial pathogens in food, the European Regulation (CE) 2073/2005 refers to ISO standards, although alternative methods are allowed if their performance has been demonstrated to be equivalent. Based on the symptoms of the human case, a set of pathogens is looked for through stepwise cultures on selective media and if relevant, the targeting of specific genes with real-time polymerase chain reactions (qPCRs) to characterize the strain. If the contaminant is successfully isolated, for some, it is characterized with Pulsed-Field Gel Electrophoresis or Multiple-Locus Variable Number Tandem Repeat Analysis for relatedness (Fratamico et al., 2016). However, in the last decade, whole genome sequencing (WGS) of the isolate has been proposed as a higher resolution alternative for the full characterization of the micro-organisms (Deng et al., 2016; Franz et al., 2016). This approach allows the detection of all genes present on the bacterial genome in just one test as well as phylogenetic analysis to link cases of food and human origin at the single nucleotide level. This resulted in recommendations from the European Centre for Disease Prevention and Control (ECDC) and the European Food Safety Authority (EFSA) to implement WGS on isolates in Europe for surveillance and outbreak investigation for a short list of priority pathogens and diseases (EFSA Panel on Biology Hazards (BIOHAZ), 2014; ECDC, 2016; Revez et al., 2017).

However, the isolation of bacteria for the conventional method is a time-consuming process that is not always straightforward nor successful (McMeekin, 2003). In that case, the outbreak investigation cannot be resolved at the microbiological level. Although qPCR-based detection methods of the food matrices, i.e., without isolation of the pathogen, can suggest the potential presence of the contaminant, it is not possible to link it back to the human cases. Sequencing methods with sufficiently high resolution that do not require isolation could solve this issue. A shotgun metagenomics approach consists in the direct sequencing of all DNA present in a sample. This gives an overview of the genomic composition of all cells in the sample, including the food source itself and the microbial community. This novel approach promises the detection of pathogens present in the sample without the need for isolation, avoiding problems linked to viable but non-culturable or difficult to isolate contaminants, and even circumventing the need for *a priori* knowledge about the causative agent (Forbes et al., 2017; Gardy and Loman, 2018). DNA sequencing may also allow, if a sufficient depth can be obtained, to have the complete genetic information about the pathogen (Gardy and Loman, 2018), to the single nucleotide polymorphism (SNP) level of accuracy. However, the challenge remains to correctly attribute each sequenced read to the appropriate strain with bioinformatics tools, in the presence of abundant host-originating reads, for a characterization of the pathogen's genome, including the determination of which (virulence, serotyping…) genes are occurring on the same genome. The choice of the DNA extraction procedure might affect the quality of the obtained DNA as well as the proportion of the species including the host and the pathogen's DNA. Some studies have previously researched the performances of several kits for metagenomics analysis on feces (Josefsen et al., 2015; Knudsen et al., 2016), but this has not yet been done for food. Another hurdle for the metagenomics analysis of food is the presence of the pathogens at very low abundances and the heterogeneity of the contamination in the food product. An enrichment of the target, already performed using the conventional microbiological methods, appears necessary. In previous metagenomics studies (Leonard et al., 2015; Hyeon et al., 2018), different enrichment durations have been tested with several selective broths, and the possibility to use a random DNA amplification to replace the natural growth of the bacteria has been proposed. Researchers have demonstrated the potential of short reads shotgun metagenomics to identify bacteria in naturally contaminated or spiked food samples to a species- or even strain-level precision (Leonard et al., 2016; Yang et al., 2016) or to characterize the pathogen by the detection of functional characteristics such as the presence of virulence genes (Walsh et al., 2017; Grützke et al., 2019). A metagenomics method has also shown its potential in feces to detect multiple pathogens in one sample (Huang et al., 2017) and even multiple strains of the same species (Singh et al., 2019), but this has not yet been achieved at the lower level of contamination observed in food. The detection of multiple pathogens, even from the same species, would represent an added value to metagenomics compared to all traditional analyses requiring isolation, for which commonly only one isolate is further characterized. The currently available studies rather stayed in the research laboratory setting. However, a method used for routine practices in reference and routine laboratories requires following the guidelines of the current regulations or achieving results with at least similar performances, with a standard protocol for sample preparation that can be applied to a range of different food matrices. Therefore, a thorough validation is necessary upon its implementation.

The application of a metagenomics workflow to the issue of foodborne outbreaks could be particularly useful to circumvent the need for isolation in conventional methods and get a faster response in case of outbreak investigation. The Shiga toxin-producing *Escherichia coli* (STEC) is a particularly challenging pathogen to analyze with such an approach as its minimal infectious dose is very low, which is defined as 10 colony-forming units (CFU) (Feng et al., 2011), and non-pathogenic *E. coli* are vastly represented in the environment and in food often associated with STEC contamination (Koutsoumanis et al., 2020). Therefore, it is difficult to differentiate various strains in a sample and infer the virulence genes to its corresponding genome, to characterize the specific *E. coli* pathotype, and hence its potential danger for human, based on the presence of specific virulence genes. STEC is a zoonotic disease that is mainly contracted through food consumption, but it is also related to animal contact, human-to-human contact, and water or soil absorption (WHO, 2015). It is a Gram-negative bacterium that is defined as a pathogen by its capability to produce one of two types of Shiga toxins, which are coded in a prophage containing the *stx1* or *stx2* genes (ISO: International Organization for standardization, 2012). Indeed, STEC bacteria share about 75% of their genome with non-pathogenic *E. coli* (Hayashi et al., 2001), and they acquire their toxicity through phages, which can also be present in food (Krüger et al., 2011). Another virulence factor of interest is the *eae* gene, which is present on the chromosome, and related to the production of intimin, a protein causing cell attachment to the intestinal wall. Using the presence of one or more of these genes, STEC are currently detected in routine laboratories through ISO/TS 13136:2012 (ISO: International Organization for standardization, 2012). STEC can cause bloody diarrhea that can lead to a hemolytic uremic syndrome (HUS) and even death. The severity of the disease can be predicted based on the subtype of the *stx1* and *stx2* genes, as well as the detection of other virulence factors such as the *eae, aaiC*, or *aggR* genes causing aggregative adherence to the intestinal mucosa of the host, or the *ehxA* gene responsible for the production of hemolysin (Ethelberg et al., 2004; Braeye et al., 2014). However, this is not taken into account in the current regulations or international methods and requires extra qPCR tests or the sequencing of (parts of) the isolate's genome. ECDC and EFSA have recommended the use of WGS to characterize this pathogen (EFSA, 2013; ECDC, 2016), but the acquisition of isolates is not always straightforward. Some recent STEC outbreaks have stressed the arduousness for an accurate source attribution due to difficulty in isolating (Robert Koch Institute, 2011). Therefore, it could strongly benefit from a metagenomics approach.

We present here a metagenomics workflow for the full characterization of STEC in food matrices using short reads sequencing. Our workflow was developed by testing different laboratory methods on minced beef meat spiked at the lowest infectious dose. Different enrichment and DNA extraction methods were tested in order to define a practical workflow that can be implemented in a routine setting. We evaluated the performances of the different sample preparations for the full outbreak-like characterization of STEC spiked in this complex food matrix by comparing the results obtained with a metagenomics analysis to the results obtained following the current official conventional methodology. Our bioinformatics workflow was set up in order to obtain the same output as expected from the conventional methods, i.e., the detection of *Escherichia* spiked in the sample, which is predicted here through taxonomic classification, and the prediction of the presence and severity of a pathogenic *E. coli* based on the detection of virulence factors in the sequenced reads. The genome of the STEC was then inferred, corresponding to obtaining an isolate's genome. It was characterized through gene detection and SNP phylogeny, in order to evaluate relatedness to other cases from human and food origin, as would be expected from routine analysis. Our analysis also went one step further by testing the selected workflow on samples of fresh goat cheese simultaneously spiked with two different STEC serotypes but possessing identical virulence genes.

3.2. Materials and methods

3.2.1. Spiked Sample Preparation

The experiments on beef were conducted with one strain of STEC selected from the collection of the Belgian National Reference Laboratory (NRL) (TIAC 1152, O157:H7, *stx1+, stx2+, eae+*). This strain was related to an outbreak in Limburg in 2012 (Braeye et al., 2014), and its genome was previously sequenced in another study (Nouws et al., 2020b). The inoculum preparation and artificial contamination of the food matrix was carried out according to Barbau-Piednoir et al. (Barbau-piednoir et al., 2018). Briefly, a STEC culture in Brain Heart Infusion (BHI) broth was diluted to obtain an OD600nm of 1, which was then diluted to 10−7 in buffered peptone water. An enumeration of 100 µl of the dilution was performed in triplicate on nutrient agar plates incubated for 18 ± 2 h at 37 °C (see count in Supplementary Materials Table S1). Organic minced beef meat was purchased at a local store (composition: 99.7% organic beef, natural aroma; nutritional values per 100 g on the package: energy: 464 kJ, total fat: 2 g, carbohydrates: 0 g, proteins: 22 g, salt: 1.1 g). A test portion of 25 g was 1/10 diluted in buffered peptone water (BPW), homogenized, and subsequently contaminated with 10 µl of the dilution 10−7, corresponding to the minimal infective dose of STEC (5–10 CFU). This artificial contamination was repeated three times (biological triplicates, representing biologically distinct samples accounting for random biological variation). One sample was not artificially contaminated (the "Blank", Bk).

The same procedure was followed on fresh organic goat cheese from raw milk purchased at a local store (composition and nutritional values were not specified on the package. Average values for goat cheese macronutrients per 100 g are proteins: 21.58 g, carbohydrates: 0.12 g, total sugars: 0.12 g, total fibers: 0 g, total fat: 29.84 g, (U.S. Department of Agriculture, 2020)) with the strains TIAC 1220 (O145:H28, *stx1+, eae+*) and TIAC 1878 (O103:H2, *stx1+, eae+*) from the Belgian NRL. The strains were spiked separately and co-spiked at a level of 5–10 CFU in a 25 g food matrix. For milk and dairy products, the diluent as recommended in ISO/TS 13136:2012 was used (ISO: International Organization for standardization, 2012): modified tryptic soy broth with the addition of acriflavin (12 mg/L) for inhibition of the growth of Gram-positive bacteria.

The samples were incubated for 24 h at 37 °C without shaking. The third biological replicate of the spiking was incubated for 16 h in the same conditions (methods D and E, Figure 3.1). After enrichment, 1 mL of the culture was centrifuged at 6000× g for 10 min, and the cell pellets were stored at −20 °C until DNA extraction. No fat layer was observed at the surface of the centrifuged beef, but a fat layer was observed after centrifugation of the goat cheese and was manually removed before DNA extraction, following Volk et al. (Volk et al., 2014).

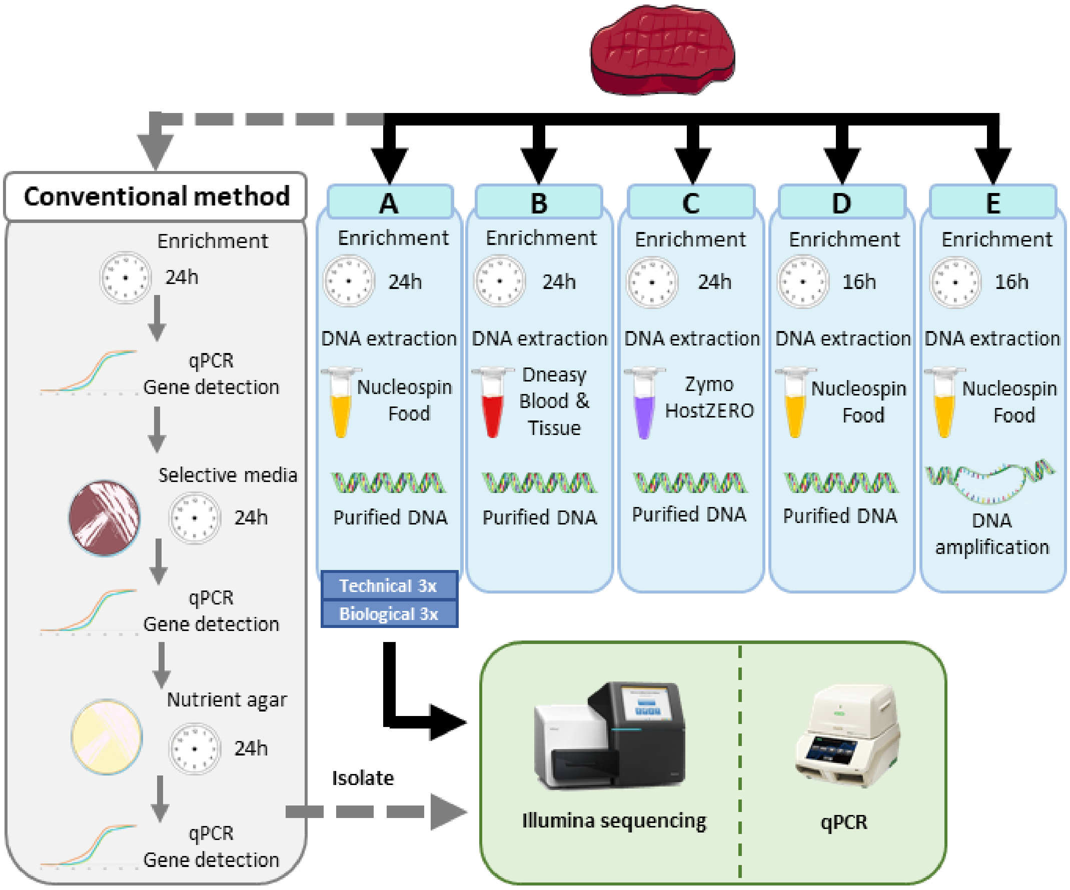


***Figure 3.1: Presentation of 5 different workflows for the preparation of metagenomics samples of spiked beef (light blue) and the conventional method for Shiga toxin-producing Escherichia coli (STEC) detection and characterization based on several steps of qPCR and isolation on selective media (ISO/TS 13136:2012, grey).** The extracted DNA (amplified or not, from metagenomics samples or isolate) is tested for quality control (DNA purity, integrity, concentration) before sequencing on the Illumina MiSeq in parallel to a qPCR check for the presence of stx genes (green).*

3.2.2. DNA Extract Preparation

Three commercial kits were used for the DNA extraction from beef samples. This was done on one of the biological replicates for the three kits: the Nucleospin Food (Macherey-Nagel, Düren, Germany, methods A, D, E, Figure 3.1), the DNeasy Blood and Tissue (Qiagen, Hilden, Germany, method B, Figure 3.1), and the HostZERO Microbial DNA kit (Zymo Research, Irvine, CA, USA, method C, Figure 3.1). DNA from all blank beef samples was extracted with the Nucleospin Food DNA extraction kit. The protocol was followed according to the manufacturer's instructions on cell pellets. The elution buffer of the DNeasy Blood and Tissue was replaced by TrisHCl 10mM. The Nucleospin Food extraction was repeated on all biological replicates. Technical triplicates were produced for the DNA extraction of the third biological replicate to account for the variability due to the extraction protocol. One DNA extract of the beef sample enriched for 16 h and extracted with the Nucleospin Food kit was amplified using phi 29 DNA polymerase (ThermoFisher scientific, Waltham, MA, USA) according to the manufacturer's instructions (method E, Figure 3.1).

Extraction blanks (extraction of water instead of the sample) were prepared for the three kits. Although no DNA could be detected using a Nanodrop 2000 (Thermo Fisher Scientific, Waltham, MA, USA), Qubit 3.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA), and 4200 TapeStation (Aglient, Santa Clara, CA, USA), they were included in metagenomics runs. The extraction blanks had very few reads, corresponding to less than 1% of the amount of reads sequenced per spiked beef samples (data not shown). The *Escherichia* genus was not detected in any of the reads from the blanks of the different kits after analysis using Kraken2 (Wood et al., 2019). Therefore, it was concluded that none of the extraction kits contained DNA that could impact the results of our analysis.

The goat cheese samples were handled according to workflow A (incubation of 24 h and DNA extraction with Nucleospin Food) as described above.

The quality and quantity of all DNA extracts were evaluated using the Nanodrop 2000 (Thermo Fisher Scientific, Waltham, MA, USA), Qubit 3.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA), and 4200 TapeStation (Aglient, Santa Clara, CA, USA).

3.2.3. Real-Time Polymerase Chain Reaction Verification

The presence of STEC with the expected virulence pattern in blank and spiked samples was verified in all DNA extracts from food matrices using qPCR for the genes *uidA, eae, stx1*, and *stx2* as described by Barbau-Piednoir et al. (Barbau-piednoir et al., 2018).

3.2.4. Validation with ISO Method

The detection of STEC in the blank and spiked samples of all biological replicates of the experiment was validated in parallel following ISO/TS 13136:2012: qPCR on the crude extract, isolation on selective media (STEC colorex, CHROMagar), confirmation of the typical colonies by qPCR, isolation on nutrient agar, and confirmation with qPCR (ISO: International

Organization for standardization, 2012) (Figure 3.1). A detailed overview of the conventional methods used for the detection and characterization of STEC in food in the Belgian NRL can be found in Nouws et al. (Nouws et al., 2020a). Then, re-isolated STEC colonies from nutrient agar plates were cultured overnight in BHI, and the DNA was extracted with the DNeasy Blood and Tissue kit according to the manufacturer's protocol. The elution buffer was replaced by TrisHCl 10 mM.

3.2.5. Next-Generation Sequencing

All DNA extracts, including isolates, were further processed with the Nextera XT library preparation kit (Illumina, San Diego, CA, USA) before sequencing on the Illumina Miseq, generating paired-end 250-bp reads with the reagent kit v3, according to the manufacturer's instructions. The samples were sequenced in three different sequencing runs, each containing libraries of 12 samples (Supplementary Materials Table S1).

3.2.6. Data Analysis

The sequence reads obtained for the isolates were further processed with the pipeline as described in Nouws et al. (Nouws et al., 2020b). The number of reads sequenced per metagenomics sample is presented in Supplementary Materials Table S2. Then, raw reads were analyzed through a bioinformatics workflow presented in Figure 3.2. The reads were trimmed using Trimmomatic version 0.38.0 operating sliding window trimming averaged on 4 bases requiring an average quality of 20 (Bolger et al., 2014) (Supplementary Materials Table S2). A taxonomic classification of the reads was conducted using Kraken2 version 2.0.7 (Wood et al., 2019) first with an in-house database of mammalian sequences containing the following genomes in order to filter out the host DNA: *Bos taurus* (GCF_000003055), *Capra hircus* (GCF_001704415), *Chlorocebus sabaeus* (GCF_000409795), *Mesocricetus auratus* (GCF_000349665), *Cavia porcellus* (GCF_000151735), *Equus caballus* (GCF_000002305), *Mus musculus* (GCF_000001635), *Rattus norvegicus* (GCF_000001895), *Ovis aries* (GCF_000298735), and *Sus scrofa* (GCF_000003025). Genomes were retrieved on 18/02/2019. Second, this was then followed by a classification on an in-house database of archaea, bacteria, fungi, human, protozoa, and viruses. This customized Kraken database was built using all available RefSeq "complete Genome" sequences of the targeted taxonomic groups downloaded from the RefSeq Genome
(ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/) on 18/02/2019 (O'Leary et al., 2016). This classification step was done using a two-step approach because a joint search against both databases requires computational resources beyond what is available for the Belgian NRL. Graphs were created on the classification results using ggplot2 in R. A gene detection was performed on all trimmed reads with SRST2 version 0.2.0 (Inouye et al., 2014) on the databases of VirulenceFinder *E. coli* and shiga-toxin genes (Joensen et al., 2014) and SerotypeFinder O type and H type (Joensen et al., 2015) as accessed in January 2020, filtering genes covered at 80% and above with a maximum divergence of 20% (results presented in

Supplementary Materials Table S2). Graphs of the depth of detection normalized to 1 million trimmed reads per sample were drawn using R and the library ComplexHeatmap (Gu et al., 2016). Strain-level metagenomics analysis was performed using Sigma (Ahn et al., 2015) following the method described by Saltykova et al. (Saltykova et al., 2020) with a database of 728 complete genomes of *Escherichia coli* from the National Center for Biotechnology Information (NCBI). Gene detection with SRST2 was performed on the classified reads corresponding to the individual *E. coli* strains with the same databases using a minimal coverage of 30% and maximal divergence of 20% as parameters. All genes detected in the strains with these parameters were considered present if they were detected with 80% coverage and identity in all reads from the sample, taking into consideration that part of the sequence might be lost in the read classification by Sigma. For phylogenetic analysis, SNP calling was carried out on the classified reads as previously described by Saltykova et al. (Saltykova et al., 2020), with *E. coli* O157:H7 str. Sakai (BA000007.2) as a reference. A matrix of SNP differences per million genomic positions covered was calculated and presented in Supplementary Materials Table S3. Maximum likelihood substitution model selection and phylogenetic tree inference were done with MEGA (Kumar et al., 2018), using the NNI (nearest-neighbor-interchange) heuristic method, keeping all informative sites and using the bootstrap method with 100 replicates as a phylogeny test. The parameters of the model selected for the construction of each tree are presented in Supplementary Materials Table S4. iTOL (Letunic and Bork, 2016) was used for the representation of the tree with percentage of reference genome covered and gene detection displayed as annotations on the side of the branches. The percentage of the reference genome represents the fraction of the genome that was suitable for the phylogenetic analysis and not the percentage of positions in the genome that were covered by reads (due to (imperfect) repeats excluded during SNP calling because of lower mapping quality). The strains used in the tree were sequenced for another study based on isolates from the National Reference Laboratory of STEC (Nouws et al., 2020b) (accession numbers: SRR10201483, SRR10201465, SRR10201452, SRR10201427, SRR10201416, SRR10201408, SRR10201398, SRR11816083, SRR11816082, SRR11816006, SRR11816065, SRR11816010, SRR11816005, SRR11816071, SRR11816075, SRR11816012, SRR11816073). The data from all metagenomics samples presented in this study can be accessed under BioProject PRJNA645436.

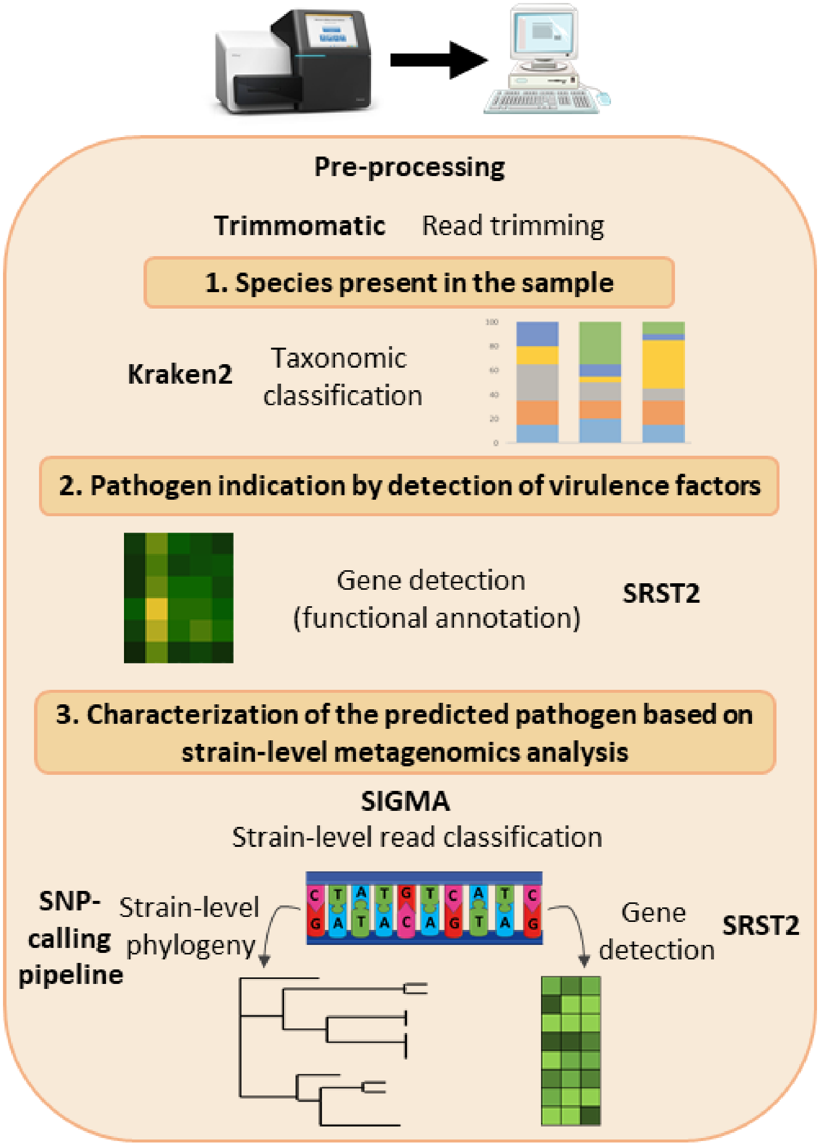

Pre-processing
Trimmomatic
Read trimming
1. Species present in the sample
Kraken2
Taxonomic classification
100
80
60
40
20
0
2. Pathogen indication by detection of virulence factors
Gene detection (functional annotation)
SRST2
3. Characterization of the predicted pathogen based on strain-level metagenomics analysis
SIGMA
Strain-level read classification
SNP-calling pipeline
Strain-level phylogeny
C T A T G T C A T C
G A T A C A G T A G
Gene detection
SRST2

Figure 3.2: Presentation of the bioinformatics analysis for the characterization of STEC in samples using a metagenomics approach. *After sequencing and pre-processing of the reads, first, the species in the sample are detected by a taxonomic classification tool (Kraken2); then, the presence of a pathogen in the sample is predicted based on the detection of virulence genes in the reads (SRST2), after which individual bacterial strains are inferred (Sigma) and characterized with gene detection (SRST2) and single nucleotide polymorphism (SNP) phylogeny (SNP-calling pipeline).*

3.3. Results

3.3.1. Testing of 5 Sample Preparation Workflows for Metagenomics Analysis Applied on Spiked Beef

Different parameters were tested in parallel for the handling of spiked beef samples as presented in Figure 3.1, to investigate which workflow would be the most appropriate and performant for a strain-level metagenomics-based outbreak investigation in reference and routine laboratories. We evaluated the performances of 5 metagenomics sample preparation workflows, which differed by the DNA extraction kit used and enrichment method (16 versus 24 h enrichment, with or without a DNA amplification step) (Materials and Methods; Figure 3.1). The outcome of metagenomics analysis of the samples was compared to the conventional methods involving isolation of the pathogen used in the National Reference laboratories, ISO/TS 13136:2012, followed by WGS of the isolate. Our metagenomics data analysis (Figure 3.2) aimed at obtaining at least the same results as the conventional methods, i.e., the detection of a STEC, serotype, and three virulence genes (*eae, stx1, stx2*), and the determination of relatedness to outbreak cases without isolation.

3.3.1.1. Comparison of the Experiment with Conventional Methods

The blank and spiked samples were tested following the conventional methods (i.e., ISO/TS 13136:2012), which are currently used in the National Reference laboratories (Supplementary Materials Table S5). All samples gave the expected results according to the spiking: detection of the stx and eae genes with qPCR in the crude extract (Supplementary Materials Table S5) and in isolates after consecutive culture steps (use of selective media and isolation of typical colonies on nutrient agar, Figure 3.1) (Nouws et al., 2020a), except in the blanks. Furthermore, sequencing of the isolates obtained from these samples with the conventional methods, an analysis recommended by EFSA but not included in the ISO, allowed the detection of the expected virulence profile and serotype in the sequencing reads. The isolate could be related to the other outbreak cases including patient’s strains in a phylogenetic analysis (see phylogenetic analysis below, Section 3.1.3, Section 3.1.4 and Section 3.1.6). These results confirmed the correct course of the spiking experiment.

3.3.1.2. Analysis of Blank Beef Samples

The absence of STEC in the food matrix was investigated using qPCR (Supplementary Materials Table S1) as well as shotgun metagenomics analysis (Figure 3.2) of the DNA extracts of the blanks. The *uidA* gene, an indicator for the presence of *E. coli*, was detected by qPCR in the DNA extracts of all blank samples, including the non-enriched blank where it was observed at a high quantification cycle (Cq 33). Some variation was present between the biological replicates of the enriched blank: the first biological replicate showed lower levels of *E. coli* (Cq 22) compared to the two others (Cq < 20).

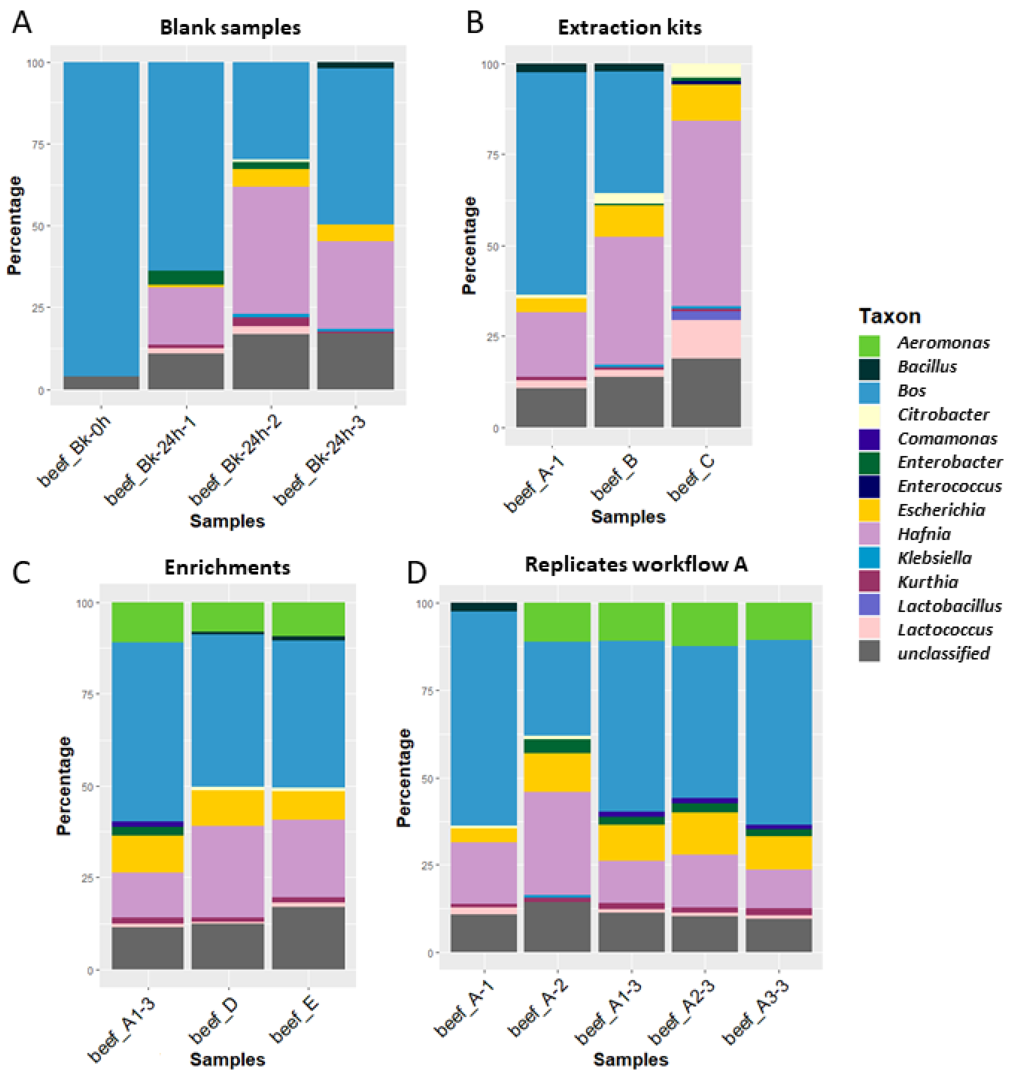


Figure 3.3: Percentages of reads classified to the genus level using Kraken2 (taxonomic classification tool) from beef samples with in-house databases of mammals, archaea, bacteria, fungi, human, protozoa, and viruses. *Light blue represents the proportion of "Bos" corresponding to beef reads. Yellow represents the presence of "Escherichia" in the sample. The reads that could not be classified to the genus level for mammals, archaea, bacteria, fungi, human, protozoa, or viruses are represented in gray. (A) Blank meat samples; Bk-0h—non-enriched blank; BK-24h—non-spiked meat sample enriched for 24 h, 1–3 biological replicates. (B) Extraction kits; workflow A—Nucleospin Food, workflow B—DNeasy Blood & Tissue and workflow C—Zymo HostZERO. (C) Enrichment times; workflow A—24 h culture enrichment, workflow D—16 h culture enrichment, workflow E—16 h culture enrichment, extraction followed by DNA amplification using phi 29 DNA polymerase; all extracted with Nucleospin Food kit. (D) Biological and technical replicates of workflow A. Small differences in the detected species shown in panels A, C, and D can be explained by the heterogeneity of the samples and biological variation, as different replicates of the experiment were used.*

Figure 3.3A presents the taxonomic distribution to genus level in each blank sample. The non-enriched blank of the first biological experiment (beef_Bk-0h) presented 96% of reads classified as *Bos*, corresponding to the host food matrix (beef meat). No bacterial species could be detected without enrichment. After enrichment, the three biological replicates of the blank (beef_Bk-24h-1, -2, and -3) harbored on average about 50% of *Bos* reads (47%, 29%, and 64% respectively), while the other reads were distributed between bacteria naturally present in the beef meat. Of these, *Escherichia* was detected in each biological replicate as one of the major taxons, indicating that the bacteria of this genus were naturally present in the beef matrix. Therefore, the sequencing results confirmed the qPCR results for the *uidA* gene obtained for the food mix crude extract analyzed with the conventional method (Supplementary Materials Table S5), and the qPCR results on the DNA extracts of this food mix (Supplementary Materials Table S1). *Escherichia* was detected with less reads in the first biological replicate (1%), confirming the qPCR observations (higher Cq, Supplementary Materials Table S1). *Hafnia* and *Kurthia* were also detected in all blanks. These species are commensal in ruminants and often detected in meat products (Davies et al., 1998; EFSA, 2016). Although *Hafnia alvei* has been described in rare cases as an opportunistic pathogen to humans, it is not considered to contribute to meat spoilage or pose a risk to human health (EFSA, 2016; Food Safety Authority of Ireland, 2019). We did not observe an issue with the integrity of the meat when the analysis was conducted (within the expiry date). Some other less represented genera varied between the samples (i.e., *Klebsiella, Enterobacter, Citrobacter, Bacillus*, and *Lactococcus*).

The genes *stx1, stx2*, and *eae* were not detected in any of the blank samples with qPCR (Supplementary Materials Table S1) nor with gene detection in the sequenced reads (Figure 3.4, Supplementary Materials Table S2). These observations, as well as the taxonomic classification, indicate that endogenous non-pathogenic *E. coli* were already present in the beef sample prior to spiking, as detected by qPCR. Accordingly, no STEC strains could be inferred for relatedness according to our bioinformatics protocol (Figure 3.2), although reads from non-pathogenic endogenous *E. coli* could be obtained

3.3.1.3. Testing of 3 DNA Extraction Kits for the Spiked Beef Samples

The DNA extraction of spiked beef enriched for 24 h was conducted with 3 different commercial kits (methods A, B, C) on the same sample (first biological replicate). The three DNA extracts tested positive for the presence of *E. coli* DNA, as indicated by the *uidA* qPCR assay, with a Cq of 20 for workflow A, 19 for workflow B, and 18 for workflow C (Supplementary Materials Table S1).

.

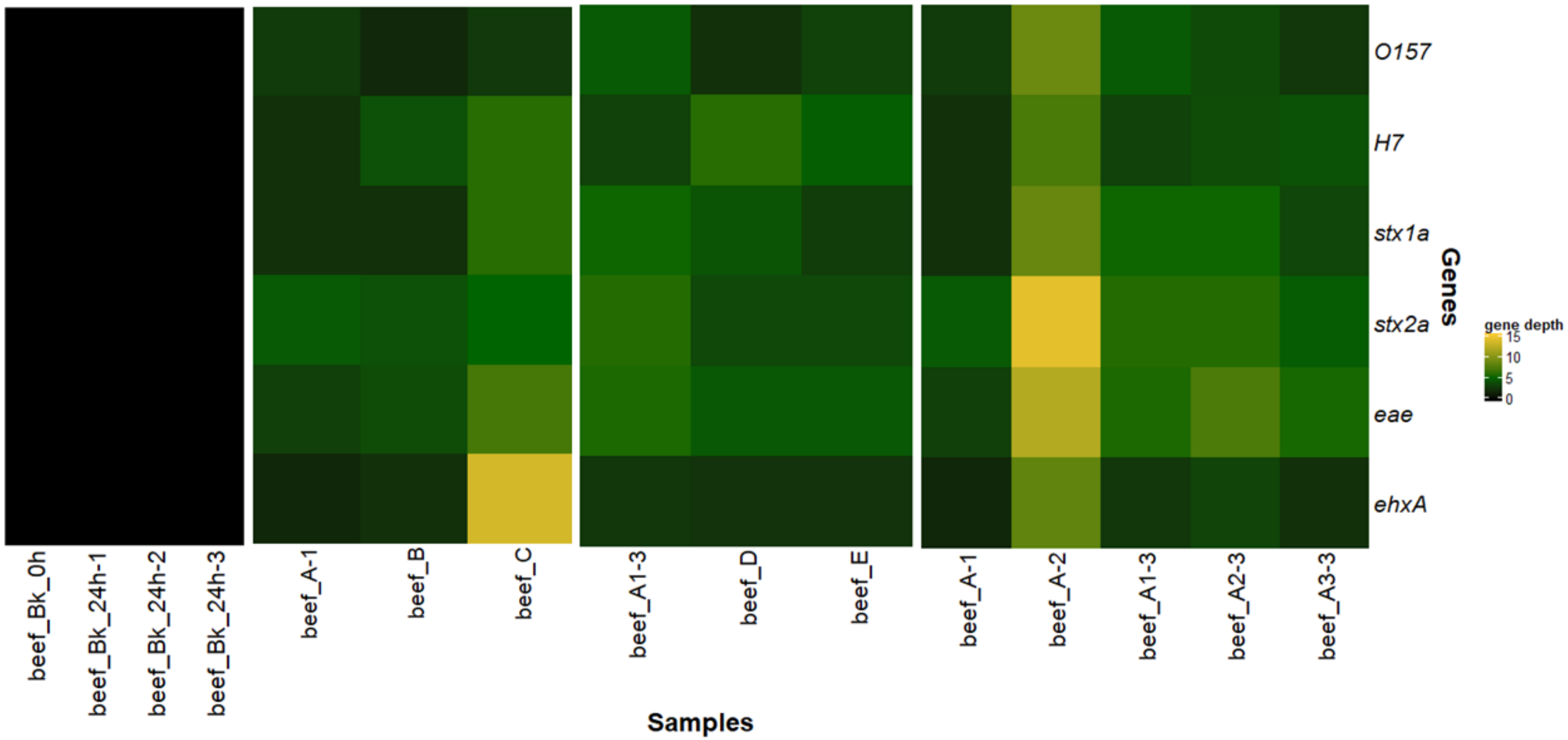


***Figure 3.4: Gene depth per million trimmed reads per sample for the detection of genes encoding for serotype O157:H7 (wzx and fliC genes) and the stx1a, stx2a, eae, and ehxA virulence genes (5 genes from ISO/TS 13136:2012 and ehxA present on plasmid pO157)** with more than 80% query coverage and 80% identity in all reads for beef samples processed with different workflows A-B-C-D-E, and in biological (A-1, A-2, A-3) and technical replicates (A1-3, A2-3, A3-3) of workflow A. Increasing depth (per million trimmed reads) is represented in shades of green to yellow according to the color gradient in the legend.*

After taxonomic classification, *Escherichia* reads could be retrieved in all DNA extracts from the spiked and blank samples (Figure 3.3B), and in higher proportions than in the corresponding blank (beef_Bk_24h-1, Figure 3.3A), as expected from the artificial addition of the STEC strain. No reads were classified as *Bos* in the sample extracted with HostZero (workflow C), demonstrating the efficiency of the kit to remove the DNA of eukaryotic cells. Meanwhile, 9.8% of the reads of the HostZero DNA extract were classified as *Escherichia*, while a very high percentage of all bacteria were classified as *Hafnia*. The use of the two other kits (workflows A and B) led to similar profiles with some variations in the percentages of the taxa. *Escherichia* was correctly detected in those DNA extracts (4% for workflow A and 8.5% for workflow B).

The serotype and the virulence (based on the detection of 6 genes) were correctly determined after extraction with the three different extraction kits (Figure 3.4 and Supplementary Materials Table S2), confirming the results of the qPCR (Supplementary Materials Table S1). Even the subtyping of the *stx1* and *stx2* virulence factors could be achieved. After normalization per million trimmed reads, *ehxA*, a gene present on the pO157 plasmid, was mapped with less reads (lower depth) than the ones present on the chromosome for workflows A and B, but with more reads for workflow C. The DNeasy Blood and Tissue kit (workflow B) had overall less reads mapping to all studied genes.

Strain-level metagenomics analysis performed on the WGS data of complex food matrices indicated that the spiked samples contained more than one *E. coli* strain. For all samples, one strain was identified as STEC by the detection of *stx* genes and was used in the phylogenetic analysis. The other strains are considered as endogenous *E. coli*. For all three DNA extraction kits, the detected STEC strain clustered with food (TIAC 1151, 1152) and human (TIAC 1169, 1165) isolates of the Limburg outbreak with 0 to 1 SNPs difference per million genomic position (Supplementary Materials Table S3) and separated from non-outbreak isolates (TIAC 1153 and TIAC 1638) (Figure 3.5 A, B) on the phylogenetic tree. All DNA extraction kits allowed the determination of the 6 genes (typing genes *wzx* and *fliC* corresponding to serotype O157:H7 and virulence genes *stx1, stx2, eae* and *ehxA*) in the reads of the STEC strain, confirming that it was the spiked strain. The subtype of the *stx* genes could also be obtained in the inferred strains. The percentage of the reference genome covered after DNA extraction of the same biological sample, presented on the side of the tree in Figure 3.5B, was higher with workflow C (HostZero extraction) and lower for workflow B (DNeasy Blood & Tissue), but all were in line with the percentages observed for isolates, indicating that a sufficient sequencing depth could be achieved for the spiked strain for a robust phylogenetic placement. The remaining *E. coli* strains detected in the metagenomics samples were also screened for the presence of virulence genes but resulted negative for the detection of *stx1, stx2, eae*, and *ehxA*, indicating that they represented the endogenous *E. coli* present on the food matrix. Therefore, these were not investigated further.

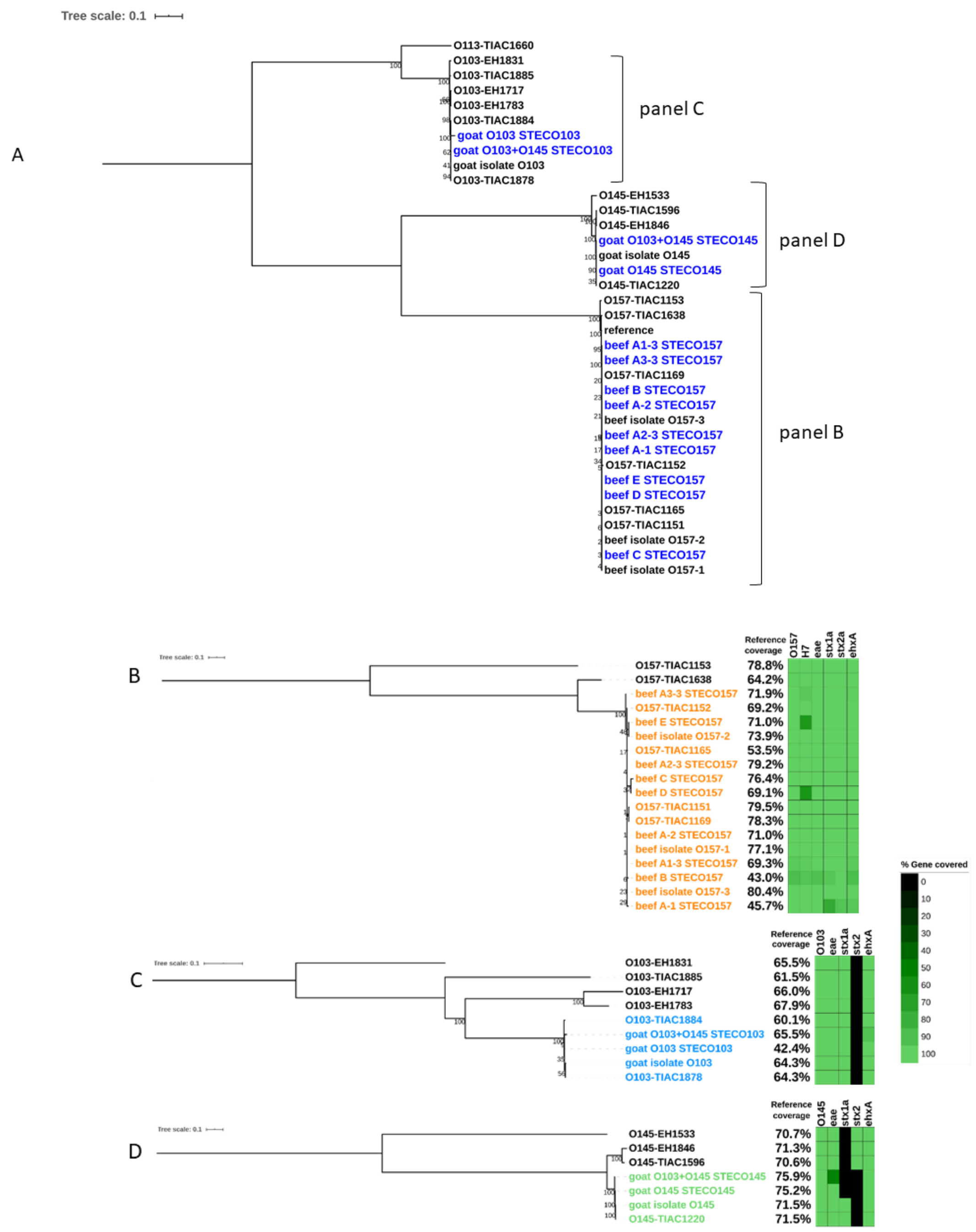


Figure 3.5: (A) SNP-based phylogenetic tree of STEC strains inferred from metagenomics samples (dark blue) and of sequenced isolates (black). *Reference: E. coli O157:H7 str. Sakai (BA000007.2). Beef/goat isolate: STEC isolate obtained after following the conventional method on the prepared spiked samples.* ***(B) Phylogenetic tree of the STEC O157 with percentage of the reference genome covered and gene detection in the strains. Orange: closely related strains from the outbreak cluster. (C) Phylogenetic tree of the STEC O103 with percentage of the reference genome covered and gene detection in the strains. Blue: closely related strains. (D) Phylogenetic tree of the STEC O145 with percentage of the reference genome covered and gene detection in the strains.*** *Green: closely related strains. The scale bar represents nucleotide substitution per 100 nucleotide site. Node values represent bootstrap support values.*

3.3.1.4. Testing of Different Enrichment Procedures

To see whether the enrichment of 24 h could be shortened, workflow A (24 h) was compared on the same sample (third biological replicate) to an enrichment of 16 h (workflow D) and an enrichment of 16 h with the same DNA extraction kit (Nucleospin Food), followed by a DNA amplification using phi 29 polymerase (workflow E).

All extracts were positive for *uidA* tested in qPCR (Supplementary Materials Table S1). All three workflows showed comparable Cq values (A and D: 20, E: 23). Sequencing and taxonomic classification confirmed the low variation between the samples (Figure 3.3C). Escherichia was detected in the three DNA preparations and represented between 8% and 10% of the reads. The amplification of the DNA (workflow E) did not qualitatively affect the distribution of the species in the sample, but resulted in a higher amount of unclassified reads. Two low abundance species (*Comamonas* and *Enterobacter*) detected in the 24 h enrichment were not detected in the 16 h enrichment samples (D and E). This can be due to growth differences during the culture and/or incubation time, which were conducted in two separate bags.

qPCR on the virulence genes of interest (Supplementary Materials Table S1) gave similar results for the sample enriched for 16 h (workflow D) and the sample enriched for 24 h (workflow A). The sample processed with workflow E (16 h of enrichment and DNA amplification) had slightly higher Cqs for all genes tested. After sequencing, the 5 genes from the ISO standard and ehxA were detected in the DNA extracted from the food samples processed with the three different workflows (Figure 3.4). The time of enrichment did not impact the gene detection, as the depth (normalized per million trimmed reads) was in the same range for the food samples enriched for 16 or 24 h (Figure 3.4). The DNA amplification (workflow E) had similar results to the sample preparations without amplification (Figure 3.4 and Supplementary Materials Table S2).

After strain-level metagenomics analysis, one STEC strain was inferred from each metagenomics sample while other endogenous *E. coli* were present in the same samples. The obtained STEC strain could be related to the isolates from food and human origin of the same outbreak for the three workflows (Figure 3.5 A, B) in a phylogenetic analysis. The obtained strain of each of the three workflows had 0 SNP difference per million genomic position with the outbreak isolates (Supplementary Materials Table S3) and shared similar SNP differences as these isolates have with the non-outbreak cases. All 5 genes from the ISO standard could be detected in the inferred genomes, as well as *ehxA*, although a lower coverage for the gene coding for type H7 was observed in the DNA extracted from the food samples enriched for 16 h (workflows D and E).

3.3.1.5. Evaluation of the Performances of the Tested Metagenomics Workflows

As elaborated above, all workflows allowed a characterization of the pathogen after enrichment, comparable to the conventional method, but without prior isolation: i.e., detection of STEC in the sample, determination of the serotype and virulence factors of interest, and retrieval of the reads corresponding to the STEC strain to perform phylogenetic tracing back to the Limburg outbreak. However, some small drawbacks were noted. Workflow B, although easy to implement for an average cost, resulted in low depths for the detection of the genes of interest and a lower coverage of the genome for the reference strain. Workflow C yielded very good results for gene detection and linkage to the outbreak isolates as well, but it is more expensive compared to the other methods. Workflow E did not show sufficient added value of the DNA amplification to pursue this additional step. Workflows A and D consisted of DNA extraction with the Nucleospin Food kit, differing in the enrichment time (24 or 16 h). Overall, workflow A showed a good performance for gene detection and strain-level metagenomics analysis with a short hands-on time and low price per sample. Additionally, 24 h of enrichment followed by a fast DNA extraction protocol seemed to be the most practical during outbreak investigation and to be used in a reference or routine laboratory setting, as it is in line with the current ISO regulation (ISO/TS 13136:2012). Therefore, workflow A was selected for further analyses.

3.3.1.6. Reproducibility of Workflow A

The reproducibility of the selected workflow A was verified using biological and technical replicates, representing the random biological variation due to the spiking and enrichment of the samples, and the variability due to the extraction protocol, respectively. After DNA extraction, the *uidA* gene was detected with similar Cq (19.34 to 20.46), as tested with qPCR (Supplementary Materials Table S1). After sequencing, the obtained reads were classified per taxon (Figure 3.3D). *Bos* accounted for approximately half of the reads in all samples, while *Kurthia*, *Hafnia*, and *Escherichia* were also present in every replicate. *Escherichia* was detected at various levels in all the spiked samples, roughly twice as much as in the corresponding blanks (Figure 3.3A). The increase of bacteria of this genus results from the addition of the STEC inoculum, introducing a new strain of this genus in the mix. The same species were detected in the three technical replicates (A1-3, A2-3, A3-3), and the difference in the species distributions was low. The biological replicates had more variation, and the first one presented the least reads classified as *Escherichia* in the blank and the spiked sample.

The genes of interest were detected in all replicates of the experiment using qPCR (Supplementary Materials Table S1) and in the sequenced reads (Figure 3.4). The gene *ehxA* was detected with a lower depth in all replicates. After normalization to a million trimmed reads, the first biological replicate of the experiment had overall lower depths for the 6 genes, while the second biological replicate showed higher amount of reads mapping to each gene of interest, but all genes could be detected, including the subtype of *stx1* and *stx2*, for all samples. This can be linked to the lower percentage of reads classified as *Escherichia* in the DNA extracted from this food sample (Figure 3.3A) as well as a higher Cq in qPCR (Supplementary Materials Table S1). The variation between technical triplicates was lower than that between the biological replicates.

Strain-level metagenomics analysis allowed the detection of a STEC strain in all replicates of the experiment. These strains clustered correctly with the isolates linked to the outbreak on the phylogenetic tree (Figure 3.5A, B) with 0 to 1 SNPs distance per million genomic positions (Supplementary Materials Table S3). In every case, the strain harbored the serotyping and virulence genes as used in ISO characterization, as well as *ehxA*, and covered about 70% of the reference genome, except for the first biological replicate that covered 45.7% of the reference genome.

3.3.2. Detection and Characterization of Two STEC Strains in Goat Cheese

The same metagenomics analysis was conducted on spiked goat cheese samples in parallel with the conventional method, as a reference. This matrix is known to be difficult to analyze due to its high fat content and complex bacterial community (Volk et al., 2014). Additionally, we increased the complexity further by spiking two other STEC serotypes (O103 and O145) harboring identical *stx* and *eae* genes separately and in a co-contamination scenario (the two strains were spiked simultaneously). Sample preparation was conducted following the selected workflow A (enrichment of 24 h and extraction using Nucleospin Food, Figure 3.1), and the same bioinformatics analysis (Figure 3.2) was applied.

3.3.3.1. Comparison of the Experiment with Conventional Method

The blank and spiked goat cheese samples were tested following the conventional methods (ISO/TS 13136:2012) to verify the spiking step. The blank goat cheese (not spiked) was negative for *uidA, eae, stx1*, and *stx2* with qPCR, and no isolate could be obtained (Supplementary Materials Table S5). All spiked samples could be completely characterized with the conventional method, including the co-spiked sample: i.e., the virulence genes were detected in the enriched food matrix (*eae, stx1*) with qPCR, the isolation of the STEC strain(s) on selective media, their isolation on nutrient agar, and the confirmation of obtaining of an STEC with qPCR was achieved. After sequencing of the obtained isolates with WGS, the expected virulence profile and serotype were obtained for both strains based on the corresponding gene detection in the sequenced reads. Then, the isolates were placed in phylogenetic trees after SNP calling (Figure 3.5A,C,D). These results were in agreement with the spiking of the samples.

3.3.3.2. Metagenomics Analysis

The DNA extracts of the goat cheese samples were first tested with qPCR (results presented in Supplementary Materials Table S1). The presence of *E. coli* was detected in the DNA extracted from all spiked samples (*uidA*, Cq of 16 to 22). It was not detected in the blank. The virulence factors *eae* and *stx1*, harbored on the genomes of the two spiked strains, were detected with a Cq of 16 in the DNA extracted from all spiked goat cheese samples (results in Supplementary Materials Table S1).

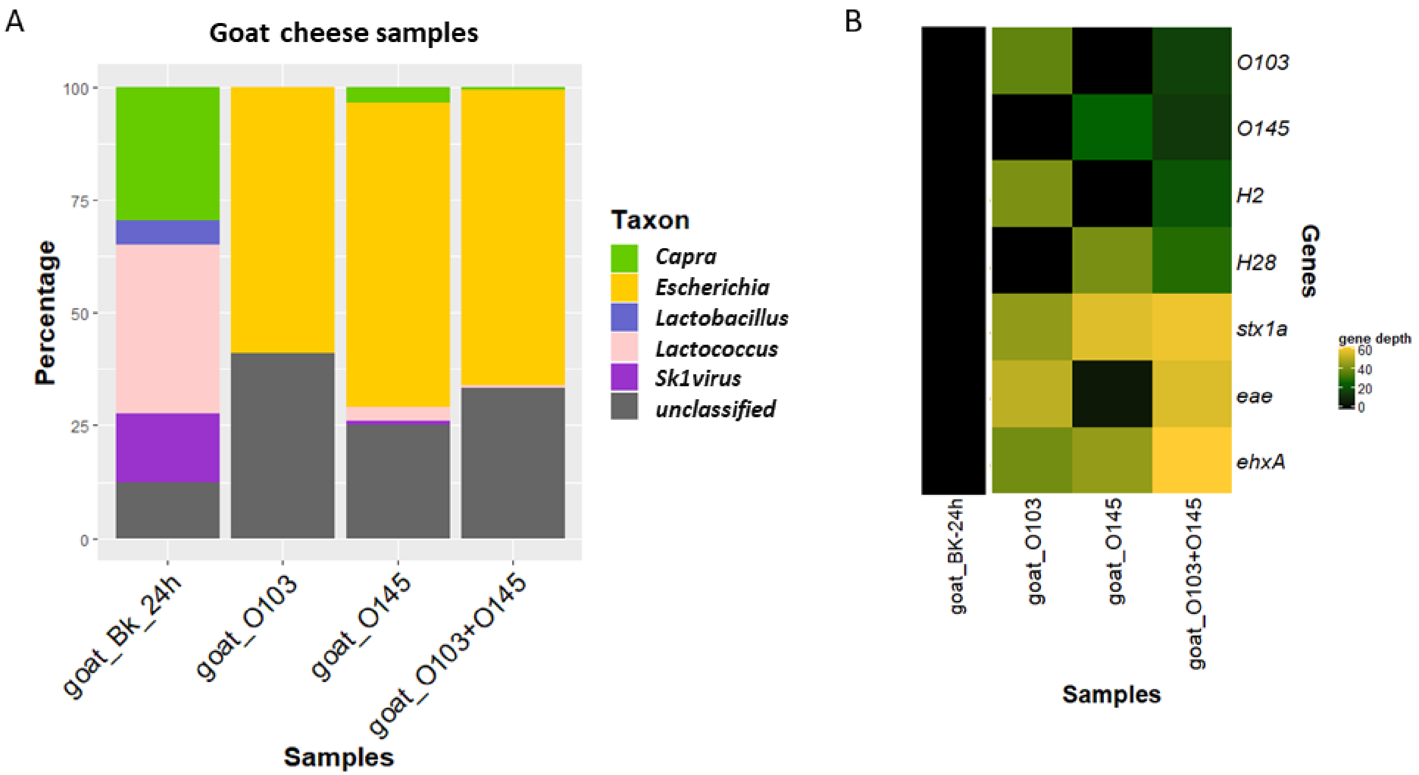


Figure 3.6: Percentages of reads classified to the genus level using Kraken2 (taxonomic classification tool) on all reads of goat cheese samples with in-house databases of mammals, archaea, bacteria, fungi, human, protozoa, and viruses. *Green represents the proportion of "Capra" corresponding to goat reads. Yellow represents the presence of "Escherichia" in the sample. The reads that could not be classified to the genus level for mammals, archaea, bacteria, fungi, human, protozoa, or viruses are represented in gray. (B) Gene depth per million trimmed reads per sample of wzx and fliC genes for the determination of types O103, O145, H2, and H28 and stx1a, eae, and ehxA virulence genes with more than 80% coverage and 80% identity in all reads of goat cheese samples. Increasing depth (per million trimmed reads) is represented in shades of green to yellow according to the color gradient in the legend. Goat_Bk_24h = Blank goat cheese enriched for 24 h. Goat_O103 = goat cheese spiked with STEC O103. Goat_O145 = goat cheese spiked with STEC O145. Goat_O103+O145 = goat cheese co- spiked with STEC O103 and STEC O145.*

After sequencing, goat DNA (Capra genus) could be detected in the blank and in small percentages in DNA extracted from the goat cheese sample spiked with STEC O145 (3%) and co-spiked with STEC O103 and STEC O145 (0.6%), but not in the sample spiked only with STEC O103 (Figure 3.6A). More reads were unclassified in that sample. The blank was also composed of *Lactobacillus*, *Lactococcus*, and SK1virus (a *Lactoccocus* virus). Only some *Lactococcus* were still detected in the spiked samples, and more than 50% of the reads were classified as *Escherichia*, which is a genus that was not naturally present in the goat cheese before the spiking.

All genes of interest that were expected to be found linked to the spiked strain could be retrieved with a high depth in the sequencing reads of the DNA extracted from the spiked goat cheese samples (see Figure 3.6B). The subtype of the s*tx1* gene could be identified without ambiguity and was identical for the two spiked strains, as expected from the analysis of the isolates using WGS. The depth or number of reads mapping to the virulence factors in the co-spiked sample was much higher than the number of reads mapping to serotyping genes in the co-spiked samples, which is as expected, because the virulence factors were to be found in the two strains, while both strains have a different serotype.

After strain-level metagenomics analysis, one strain was detected for each of the single spiked samples, while two were detected for the co-spiked sample. The strains were identified as STEC O103 and STEC O145 based on the detection of the serotyping alleles. These were placed in a phylogenetic tree in proximity of the genomes of sequenced isolates, including the spiked isolates (Figure 3.5A). Figure 3.5C depicts only the cluster of the STEC O103. The strains obtained from metagenomics samples and the corresponding spiked isolate were separated from the other sequenced isolates. The inferred STEC strain from the co-spiked sample had 2 SNPs difference per million genomic positions to the corresponding STEC O103 isolate, while the strain from the sample spiked only with STEC O103 had 1 SNP difference to this isolate (Supplementary Materials Table S3). The O-type, *stx1, eae*, and *ehxA* genes found in the isolate could be detected in the strains from the metagenomics samples. The spiked strain could be fully characterized using the reads obtained from the single as well as the co-spiked metagenomics samples. Interestingly, two isolates from the National Reference Laboratory (TIAC 1884 and TIAC 1878) were placed together in this cluster, with 1 SNP difference per million genomic positions (Supplementary Materials Table S2). These isolates come from food samples received at the same time period but that were not tested for relatedness at that time. Figure 3.5D details the cluster of the STEC O145. The inferred strains from metagenomics samples could be linked to the corresponding STEC O145 isolate with 2 and 0 SNPs difference per million genomic positions for the co-spiked and single spiked samples, respectively, and they could be separated from other sequenced isolates with the same range of SNPs distance as observed for TIAC 1220 to these other cases (Supplementary Materials Table S3). However, gene detection did not allow full characterization of the STEC O145, as *stx1* could not be detected in the inferred genome with the set parameters, although it was detected with a high depth in all reads from the samples (Figure 3.6B). However, *eae*, *ehxA*, and the serotype could be correctly detected, and the strain was placed in a cluster with the corresponding STEC strain, which indicates that there is a strong possibility that this strain is indeed a STEC. The reads containing the *stx1* gene were mapped to a separate sequence from the database when performing the Sigma workflow, which was not included in the reads corresponding to the STEC O145 strain.

3.4. Discussion

To rapidly confine foodborne outbreaks, it is important to be able to identify and characterize the food pathogen and to link it with the patient's strain, in order to take appropriate measures to prevent further spreading as quickly as possible. Conventional microbiological detection methods are based on culturing steps to obtain an isolate of the pathogen. This increases the turnaround time of the analysis and is not always successful. Besides, these methods are based on low resolution technologies such as qPCR. Recently, with the advance of WGS, the resolution has been significantly increased, although an isolate is still needed. With shotgun metagenomics, this issue would be resolved, as all DNA of the sample is sequenced, and prior isolation is not needed. However, the difficulty lies then in the correct characterization of the pathogen and the subsequent source tracking based on the metagenomics reads, i.e., a mix of everything in the sample. There is a need to disentangle the reads of pathogens present in the food sample before being able to characterize these and to link these to the patient's isolates. Moreover, the method should be adapted to an application in national reference and routine laboratories. Our study tested the performances of 5 sample preparation workflows for a short read shotgun metagenomics analysis of contaminated foods. The workflows were defined to be as close as possible to the standard methods currently used in many (reference) laboratories in Europe (isolation according to ISO/TS 13136:2012, followed by relatedness analysis in case of an outbreak). Therefore, we worked with very low loads of contamination (<10 CFU for 25 g of food) and enrichment media that fit the requirements of the ISO standard. Moreover, as other studies previously highlighted the need for an enrichment of the samples to obtain a high resolution in the analysis, we tested enrichment times approaching 24 h. Our results proved the feasibility of a metagenomics method to obtain the same information as the conventional methods in a manner that is relatively easily applicable in laboratories, and this in a shorter time period, as no isolation is needed. Time is a crucial factor during a foodborne outbreak investigation. This analysis was performed for samples containing multiple strains of *E. coli* and even several different strains of STEC. We also managed, for the first time, to link individual STEC strains from different food matrices containing multiple (including endogenous) *E. coli* strains to genomes from human cases, which is essential in resolving an outbreak.

In order to evaluate the possibility to implement our method as a new approach applicable in routine and demonstrate equal performance, we systematically compared the results obtained with metagenomics to the information collected using conventional methods. Therefore, our bioinformatics analysis was targeted at obtaining information comparable to that obtained with conventional methods. Such an approach has not yet been followed in other studies that were more focused on the research aspect for proof of concept. The possible presence of an STEC in the food sample was first evaluated with a taxonomic classification tool and the screening for virulence genes through all sequencing reads. This step corresponds to the screening stage in the conventional method (qPCR of specific virulence genes on the crude extract of the enriched test portion) and allows predicting if a potential pathogen is present in the sample. Although more sensitive tools exist (Scheuch et al., 2015), the taxonomic classification tool (Kraken2) was chosen for its fast execution (results in a few minutes) (Wood et al., 2019), which is appreciated for fast outbreak resolutions. Then, a strain-level classification of the metagenomics reads corresponding to the isolation of the strain was conducted. The obtained strain is characterized through gene detection and SNP phylogeny, which is equivalent to qPCR, followed by PFGE for relatedness or WGS analysis on an isolate in routine. However, the information obtained with the metagenomics analysis exceeded the one obtained with the conventional workflow. Indeed, information that is not requested in the scope of the ISO standard was also obtained for three different STEC serotypes tested: the metagenomics method was capable of distinguishing the subtype of the *stx1* and *stx2* genes but also to detect the gene *ehxA*, which are all recognized as markers for the severity of the disease but not included in the regulations (De Rauw et al., 2019b). Shotgun metagenomics, such as whole genome sequencing, allows obtaining an overview of the complete genome of the organisms present in a sample, therefore giving access to all genes of interest present on this genome. In further studies, other genes of interest including antimicrobial resistance genes but also other virulence genes recognized for their importance in other outbreaks such as *aaiC* or *aggR* (not present in the spiked strains) could also be investigated. Obtaining the complete genome of the STEC strain also allows the detection of any serotype, while only O26, O103, O111, O145, and O157 are currently looked for with the current methods.

The use of molecular methods in routine for food monitoring requires a DNA extraction protocol that is easy to implement with low costs and reproducible results. Therefore, we tested three commercial DNA extraction kits and two different food matrices, including one that is recognized as difficult due its higher fat content. All methods performed sufficiently well to detect and characterize the pathogenic strain in the food matrix and to link it to other outbreak cases from food and human origin. However, as previously shown, the choice of the kit can have a minor impact on the results obtained with a metagenomics study (Josefsen et al., 2015; Knudsen et al., 2016). Indeed, it can cause a variation in the distributions and even the detection of genera in the sequencing reads of the same sample. However, the presence of some taxa could also be explained by carry-over or index misidentification due to the used sequencing technology (Kircher et al., 2012) or performance of the taxonomy classification tool (Wood et al., 2019). It has also been previously described that commercial DNA extraction kits can have different performances for the extraction of plasmid DNA (Delaney et al., 2018), as observed in our study for the gene *ehxA*. As Nucleospin Food (workflow A) had good results for a low price and hands-on time and proved to be reproducible, it was selected for further experiments.

ISO/TS 13136:2012 demands an enrichment of the food matrix before the start of the analysis in order to be able to detect low levels of STEC. Moreover, implementation in a routine laboratory setting requires that the timing of the enrichment is practical for the technicians' work schedule and allows a fast analysis, which is especially important in case of outbreak investigation. Previous studies have stressed the importance of an enrichment for metagenomics analysis of food samples and have focused on a shortening of the incubation time by using selective media or antibiotics (Leonard et al., 2015; Hyeon et al., 2018). In our study, the enrichment culture was conducted in the broths recommended in the ISO standard: the goat cheese was enriched in modified tryptic soy broth with acriflavin, and a non-selective broth (buffered peptone water) was used for the analysis of spiked beef, as it is recommended for stressed cells. The use of selective media has been shown to induce the identification of only certain serogroups of STEC (Brusa et al., 2016). We analyzed three different serotypes of STEC and noted no difference in the performances of our workflow. Two enrichment times were tested in our study (i.e., 24 h and 16 h). The information obtained after 16 h of enrichment, with the applied sequencing conditions, was already sufficient for outbreak investigation purposes with similar results to those obtained with the conventional methods, and we were able to conduct this analysis to the SNP level with the presence of endogenous *E. coli* in the food matrix. A DNA amplification, after 16 h of enrichment in non-selective broth, was not considered as an added value in the protocol, in contrast to what was previously reported after 12 h of incubation in selective broth followed by selective immune-based enrichment of the pathogen's DNA (Hyeon et al., 2018). Although a shorter enrichment time can be considered, we chose to pursue our study with an enrichment of 24 h, corresponding to what is currently performed in routine and is also recommended for the analysis of STEC under stress conditions (Jasson et al., 2009).

In routine, the isolation of a strain almost automatically involves the characterization of a single pathogenic strain present in the food matrix for relatedness. However, previous studies have shown the prevalence of co-contaminations, including several different pathogens or multiple strains of the same species (Kinnula et al., 2018; Somerville et al., 2018). Therefore, it is important to develop a method that can characterize all pathogenic strains in the food vehicle for outbreak investigation. In our study, we managed to characterize STEC in food in the presence of endogenous non-pathogenic *E. coli* and in a sample co-contaminated with two different strains (serotypes O103 and O145), which is a level of analysis that was not achieved in previous studies (Leonard et al., 2016). We were able to extract the pathogen's reads and link them back to human cases using phylogenetics analysis, starting from very low levels of inoculum. This information might not be obtained for two separate strains in a routine setting, as only one STEC (when the same profile of virulence genes of single colonies is obtained by qPCR) is usually characterized for relatedness after isolation. Moreover, as metagenomics is a "pathogen-agnostic" approach, it allows the analysis of a food product without the need of *a priori* knowledge on the pathogen or the number of strains that might be present. The conventional methods of analysis of foodborne outbreak samples rely on symptom-based screening for a pathogen. In one out of four foodborne outbreaks, the causative agent cannot be identified with the current method (EFSA, 2019a). This can be caused by the absence of leftover suspect food, by a limited quantity of leftover food impeding the ability to conduct several conventional tests, or due to the difficulty to obtain an isolate to characterize further. The implementation of a metagenomics approach would allow a complete screening of possible pathogens in one test with a limited amount of sample. This might lead to interesting results as shown with the detection of other species in the enriched beef and goat, including Hafnia, which is a rare possible source of infection to humans, for example in immunocompromised patients. Although this study has focused on STEC as a bacterial foodborne pathogen, we believe that the same approach could be applied to others with only minor modifications, such as the database used for read classification and gene detection.

The possibility to implement new approaches in routine settings will also depend on the cost of the analysis per sample. Metagenomics studies still represent a high investment for laboratories. Yet, metagenomics provides access to much more information at once than conventional tests, which, if all conducted in parallel, would also become expensive, in addition to requiring a large portion of sample and therefore risking missing the causing agent if only a small amount is available (which is common for leftovers from a suspect meal). The cost is primarily linked to the low number of metagenomics samples sequenced in one run, although the price of sequencing has dropped significantly in the last few years. The amount of samples sequenced might vary depending on the desired depth, and therefore, the level of information obtained per sample could be improved by sequencing at a higher depth, but then also at a higher analysis cost. The necessary depth could also be evaluated from the qPCR result of specific markers of the pathogen in the food matrix, if it is known. However, although our metagenomics and qPCR results seemed to agree, other studies have shown that qPCR results are not directly linked to metagenomics outcome (Andersen et al., 2017). Moreover, in case of an outbreak, the time to wait for the accumulation of sufficient samples to start a run can be an obstacle for a fast response, although metagenomics runs require fewer samples than the WGS of isolates before the run is complete. To reduce the number of samples in a full run, new options with lower output such as the Flongle flow cell from Oxford Nanopore Technologies still have to be investigated and might prove cost-effective. The use of a long reads sequencing technology could offer additional advantages such as reducing the turn-around time by allowing real-time analysis and preventing bleed-through by sequencing one sample per flow cell. It might also improve the reconstruction of the genomes and the species detection by bringing bigger pieces to the puzzle (Höper et al., 2016). However, the error rate of Oxford Nanopore Technologies is still relatively high and might impact the level of details obtained in the analysis, as observed by Hyeon et al. (Hyeon et al., 2018). Another important drawback for the implementation of metagenomics in routine is the need for adapted bioinformatics pipelines (Carleton et al., 2019). This has been improving in the last years with the development of new specialized tools that can be proposed in workflows such as the one presented in this study. In the future, this workflow should be implemented as a user-friendly analysis pipeline to be executed in a routine setting. This will be worthwhile, as metagenomics approaches are now increasingly being explored, and studies such as this one prove their applicability for routine laboratories.

Interestingly, our analysis was also able to detect an unanticipated link between two isolates from the Belgian NRL received during the same time period. Relatedness and typing analyses represents extra tests and therefore extra costs, and these are not performed on a standard basis for all food isolates outside outbreak investigations in a reference or routine laboratory, while it could be achieved for every sample when performing metagenomics. This highlights the added value of the whole genome sequencing of pathogens in food samples, and by extension even from environmental samples. Importantly, this will also contribute to the creation and use of a shared database of whole genome sequences, including genomes of contaminants from human origin, in order to rapidly detect relationships between linked cases. This would allow to rapidly trace back the source of a contamination, similarly to what is being done using the genomic tracking tool GenomeTrakr (Timme et al., 2018). The addition of pathogen whole genome sequences into a database could also improve our data analysis method, as Sigma, the strain-level inference tool used, is based on the use of reference genomes, for which 728 complete *E. coli* complete genomes were available in NCBI at the time of the analysis. The acquisition of circulating STEC genomes could help for the detection of strains less common in public databases such as the STEC O145 presented in this study, for which a virulence gene was missing after genome inference, although it could be detected with very high depth before the strain-level acquisition analysis. Although infrequently sequenced, STEC O145 is one of the top six most common non-O157 serotypes associated to human diseases (Carter et al., 2016), and it has previously been linked to a multi-strain outbreak in Belgium (De Schrijver et al., 2008).

In conclusion, we presented a metagenomics method developed to be as close as possible to the actual ISO standard, but without requiring isolation. This study proved the applicability of metagenomics as a valid alternative to the standard protocols that are currently used in reference laboratories with a strain-level acquisition of reads replacing the isolation. We showed that this method can equal and even surpass the information that can be obtained with the conventional workflow, in one single test, allowing access to information on all genes in the DNA of the pathogen studied and the resolution of outbreaks by linking human cases to strains from food samples. However, the cost of the method, still high, might at first impose a rational use of the approach. The metagenomics method described in this study can be used as a faster alternative when urgent results are necessary, in particular in the case of outbreaks, or as an alternative to ISO for samples in which the isolate could not be obtained. It is also suitable to study emerging strains or pathogens such as the O104 strain from an international outbreak from German origin in 2011 (Cheung et al., 2011) and would even provide the necessary sequence information to design a conventional method allowing the detection of the same strain from food for other laboratories that do not have the capacity to invest in a metagenomics approach. Moreover, the ability to discriminate and characterize several strains in case of multi-strain outbreaks is not yet covered in current procedures in routine but, as presented in this work, it can be achieved by following a metagenomics approach. New technologies allowing a metagenomics analysis at a lower cost and in an even shorter time-frame are yet to be explored further for a facilitated implementation in routine. This will only be feasible if guidelines are adapted to fit the methods that are being developed for public health and food chain safety needs. The possibility of applying whole genome sequencing and metagenomics for outbreak investigation, source attribution, and risk assessment of foodborne microorganisms has now been assessed by the EFSA (EFSA, 2019b), demonstrating the initiation of a reflection on future regulations in this matter. Studies such as ours can contribute to convincing the policy makers to adopt these new methods into practical procedures that may be applied in reference and routine laboratories in the near future.

Acknowledgements

We would like to thank the technicians from the National Reference Laboratory of STEC (Sciensano) for their help to implement the standard method in this work. We are grateful to all the time and efforts of Stefan Hoffman, Maud Delvoye and Els Vandermassen from Transversal Activities in Applied Genomics (Sciensano) for the set-up of the metagenomics sequencing runs.

Supplementary data

The following are available online at https://www.mdpi.com/2076-2607/8/8/1191/s1, Table S1: Description of the samples and qPCR result on the DNA extract, Table S2: Gene detection in all metagenomics samples using SRST2, Table S3: SNP distances matrix (per million genomic positions) for STEC phylogenetic tree, Table S4: Model selection and parameters for all trees (Figure 3.5), Table S5: qPCR results on the enriched food samples and gene detection (SRST2) performed on the isolates obtained from the samples.

CHAPTER 4
Application of a strain-level shotgun metagenomics approach on food samples: resolution of the source of a *Salmonella* food-borne outbreak

This chapter was adapted from the previously published manuscript:
Buytaers, Florence E., Assia Saltykova, Wesley Mattheus, Bavo Verhaegen, Nancy H. C. Roosens, Kevin Vanneste, Valeska Laisnez, Naïma Hammami, Brigitte Pochet, Vera Cantaert, Kathleen Marchal, Sarah Denayer*, Sigrid C.J. De Keersmaecker*. 2021. “Application of a Strain-Level Shotgun Metagenomics Approach on Food Samples: Resolution of the Source of a *Salmonella* Food-Borne Outbreak.” Microbial Genomics 7 (4). https://doi.org/10.1099/mgen.0.000547. (* equal contribution)

Authors’ contributions:

F. E. Buytaers was in charge of the conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing of the original draft and visualization. A. Saltykova helped for the software, formal analysis, data curation, review and editing. W. Mattheus and B. Verhaegen participated in the validation, investigation, resources, review and editing. N. H. C. Roosens was involved with the resources, funding, review and editing. K. Vanneste participated for the software, validation; resources, review and editing. V. Laisnez , N. Hammami, B. Pochet and V. Cantaert helped for the investigation, resources, review and editing. K. Marchal took part in the conceptualization, supervision, review and editing. S. Denayer was involved with the conceptualization, methodology, validation, investigation, resources and writing of the original draft. S. C. J. De Keersmaecker contributed to the conceptualization, methodology, validation, formal analysis, writing of the original draft, supervision, project administration and funding:

Abstract:

Food-borne outbreak investigation currently relies on the time-consuming and challenging bacterial isolation from food, to be able to link food-derived strains to more easily obtained isolates from infected people. When no food isolate can be obtained, the source of the outbreak cannot be unambiguously determined. Shotgun metagenomics approaches applied to the food samples could circumvent this need for isolation from the suspected source, but require downstream strain-level data analysis to be able to accurately link to the human isolate. Until now, this approach has not yet been applied outside research settings to analyse real food-borne outbreak samples. In September 2019, a *Salmonella* outbreak occurred in a hotel school in Bruges, Belgium, affecting over 200 students and teachers. Following standard procedures, the Belgian National Reference Center for human salmonellosis and the National Reference Laboratory for *Salmonella* in food and feed used conventional analysis based on isolation, serotyping and MLVA (multilocus variable number tandem repeat analysis) comparison, followed by whole-genome sequencing, to confirm the source of the contamination over 2 weeks after receipt of the sample, which was freshly prepared tartar sauce in a meal cooked at the school. Our team used this outbreak as a case study to deliver a proof of concept for a short-read strain-level shotgun metagenomics approach for source tracking. We received two suspect food samples: the full meal and some freshly made tartar sauce served with this meal, requiring the use of raw eggs. After analysis, we could prove, without isolation, that *Salmonella* was present in both samples, and we obtained an inferred genome of a *Salmonella enterica* subsp. *enterica* serovar Enteritidis that could be linked back to the human isolates of the outbreak in a phylogenetic tree. These metagenomics-derived outbreak strains were separated from sporadic cases as well as from another outbreak circulating in Europe at the same time period. This is, to our knowledge, the first *Salmonella* food-borne outbreak investigation uniquely linking the food source using a metagenomics approach and this in a fast time frame.

4.1. Introduction

The detection and characterization of pathogens in food aims at avoiding contamination of consumers if carried out as a continuous screening, but also at putting an end to epidemics when consumers have already been infected. According to European Union legislation, typically the analysis of a suspect food sample involved in a food-borne outbreak includes an attempt at obtaining an isolate of the micro-organism, most often by the official control laboratories, such as the National Reference Laboratory (NRL), to further characterize it, e.g. by real-time PCR (qPCR) or whole-genome sequencing (WGS) (Naravaneni and Jamil, 2005; UE, 2005; ECDC and EFSA, 2019). To unambiguously identify the source of the outbreak, the food contaminant also has to be uniquely linked to the pathogens usually obtained from human cases by the National Reference Center (NRC). This strengthens the assumption on the food source based on epidemiological studies only. However, isolation from food samples is not straightforward nor always successful, as opposed to the human samples, which typically contain higher loads of the pathogen. In these cases, the relatedness to the human isolates cannot be obtained and the outbreak is never resolved to its food source. Indeed, the European Food Safety Authority (EFSA) reported that the causative agent was unknown in 23.8 % of outbreaks that occurred in 2018 (EFSA, 2019a; Sala et al., 2020). In some cases, the wrong foodstuff can even be blamed, leading to huge economic losses in the sector (European Commission, 2011). A novel approach, i.e. shotgun metagenomics, has been investigated in recent years in an attempt to characterize the pathogen but without the need to isolate it from the food matrix (Höper et al., 2016; Kovac et al., 2017; Carleton et al., 2019); therefore, in a possibly shorter time frame and, most importantly, increasing the chance of finding the source of the outbreak. EFSA recently published an opinion on the use of WGS and metagenomics for outbreak investigation, confirming the possibility for typing and source attribution from shotgun metagenomics data, in particular if a draft reconstructed genome of the pathogen at the strain-level can be obtained (EFSA, 2019b). Until now, only a few studies have investigated the possibility of achieving strain-level characterization for pathogens in food samples; however, these did not link strains obtained from the food samples to isolates from the human cases, a prerequisite for the trace back of the outbreak (Leonard et al., 2015, 2016; Yang et al., 2016; Walsh et al., 2017). We have previously developed such a metagenomics approach to be implemented for food-borne outbreak investigations (Buytaers et al., 2020; Saltykova et al., 2020) using artificially contaminated samples, targeting the Shiga toxin-producing *Escherichia coli* (STEC), and we were able to link it back to isolates from humans. This method has, however, not yet been implemented for another pathogen or during a real outbreak.

Among food-borne outbreaks occurring in Europe, food contaminations due to *Salmonella* are the second most commonly reported cause of gastrointestinal infections (EFSA, 2019a). Salmonelloses are caused by thousands of different serovars, of which *Salmonella enterica* subsp. *enterica* serovar Enteritidis accounts for over 40 % of all infections for which the serovar has been identified. They are most often related to eggs and have been associated with a high proportion of food-borne outbreaks, due to the use of the raw product in several food preparations (EFSA, 2019a). The standard protocol for analysing food products potentially contaminated with *Salmonella* according to European Union legislation is to isolate the pathogen through several enrichment and plating steps (ISO 6579 : 2017 (ISO: International Organization for standardization, 2017)). The isolated strain is then characterized through biochemical and/or serological testing, as well as multilocus variable number tandem repeat analysis (MLVA) to infer phylogeny against a well-characterized background. However, EFSA has now recommended WGS of *Salmonella* isolates, particularly when linked to outbreaks (EFSA, 2014b). WGS offers the possibility to study the full genome of the isolate, including potential virulence and antimicrobial-resistance (AMR) genes (EFSA, 2014b; Ellington et al., 2017). It also allows the highest level of precision in relatedness studies based on SNP differences between strains, and allows sporadic bacteria to be distinguished from persistent bacteria in a food-production environment (Franz et al., 2016; Tang et al., 2019). Using metagenomics, *Salmonella* has thus far only been characterized in faeces (Loman et al., 2015; Huang et al., 2017) or in food after selective concentration of *Salmonella* genomic DNA by immunomagnetic separation (Hyeon et al., 2018). However, food samples contaminated with this species have not yet been tested with an open metagenomics approach in the scope of a real outbreak.

From September 5th 2019 until September 14th 2019, over 200 students and teachers at a hotel and tourism school in Belgium suffered from food poisoning, with symptoms such as abdominal pain, headache, diarrhoea and fever (Centre National de Référence Salmonella & Shigella, 2020; Sciensano, 2020). The outbreak was thoroughly investigated by the local authorities [regional health agency Zorg en Gezondheid, the Federal Agency for the Security of the Food Chain (FASFC) and the NRL (food and feed) and NRC (human)]. Laboratory analyses were conducted on 65 samples obtained from food leftovers and kitchen surfaces, as well as isolates from infected patients. This resulted in the identification of the contamination as being *S. enterica* subsp. *enterica* serovar Enteritidis, found in a meal prepared on September 5th 2019 by students and served in the school restaurant. The meal consisted of fish sticks with mashed potatoes and freshly made tartar sauce. After WGS of isolates from food and human origins, the source of the contamination was established as being the sauce, prepared with raw eggs (AFSCA, 2019; Centre National de Référence Salmonella & Shigella, 2020; Sciensano, 2020). A rare MLVA profile, i.e. 3-12-5-5-1, was determined for the human and food isolates by the NRC (Centre National de Référence Salmonella & Shigella, 2020). After disinfection of the kitchen and kitchen equipment, *Salmonella* was not detected anymore in environmental samples and no new cases were recorded. The outbreak was reported through the European Epidemic Intelligence Information System (EPIS) ('Urgent Inquiry' UI-608) and the Rapid Alert System for Food and Feed (RASFF, 2020.3675) and allowed the tracing of this outbreak back to an egg-producing farm in Spain, considered as the source of the contamination (Centre National de Référence Salmonella & Shigella, 2020; Sciensano, 2020). At the same time period (ongoing since 2016), another outbreak was circulating in Europe and was linked to eggs of Polish origin. However, this strain of *S. enterica* subsp. *enterica* serovar

Enteritidis was distinct from the isolates from the hotel-school outbreak and was characterized with MLVA profiles 2-9-7-3-2, 2-9-6-3-2, 2-9-10-3-2, 2-10-6-3-2, 2-10-8-3-2 or 2-11-8-3-2 (Pijnacker et al., 2019; ECDC and EFSA, 2020).

As this was an ideal case study to apply our previously developed strain-level metagenomics approach on contaminated food samples to be used during a food-borne outbreak, we received from the Belgian NRL, in parallel to the conventional investigation, two samples that were positive for *S. enterica* Enteritidis and linked to the hotel-school outbreak. Both samples were processed with a metagenomics workflow described previously (Buytaers et al., 2020). After short-read sequencing, we conducted data analysis in order to infer the pathogenic strain's genome, characterize it and link it back to the human isolates to resolve the outbreak. The food strain obtained from metagenomics reads was included in a SNP-level phylogenetic tree containing human and food isolates from the hotel-school outbreak, as well as strains related to another outbreak circulating in Europe during the same time period (Pijnacker et al., 2019; ECDC and EFSA, 2020) and other sporadic strains that occurred in Belgium in 2019. The time of analysis of such a shotgun metagenomics approach was then compared to the time necessary to elucidate this outbreak with food isolates' data.

4.2. Materials and methods

4.2.1. Sample preparation

Two aliquots of cultured food samples (i.e. a mixture of the meal components and the sauce as a separate component) linked to the outbreak were received from the NRL after a first non-selective enrichment according to ISO 6579 (ISO: International Organization for standardization, 2017) (i.e. 25 g foodstuff was mixed with 225 ml buffered peptone water and incubated for 18±2 h at 37±1 °C). The sample dish was an aluminium tray with three compartments, one for each component (mashed potatoes, fish stick, tartar sauce). The tartar sauce was tested separately as well, after confirmation that it was the probable source of the contamination. The food enrichments had been tested for the presence of *Salmonella* prior to their selection for this study, using the iQ-Check *Salmonella* II PCR detection kit (Bio-Rad) according to the manufacturer's instructions, and showed positive results (Cq of 18 and 17, respectively) as opposed to the blanks and to other samples collected in the school during the investigation. Aliquots of 4–15 ml of the two cultured food samples were stored in the fridge until metagenomics DNA extraction was carried out.

4.2.2. DNA extraction and qPCR

The sample preparation was carried out according to Buytaers et al. (Buytaers et al., 2020). Briefly, 1 ml of the aliquots was centrifuged at 6000 g for 10 min and the cell pellets were used for DNA extraction using a Nucleospin food kit (Macherey-Nagel). In order to confirm the presence of the contaminant (*Salmonella*) in the DNA extracts, a qPCR was performed for the genes *invA* and *rpoD*, according to Barbau-Piednoir et al. (Barbau-Piednoir et al., 2013).

4.2.3. Shotgun metagenomics sequencing

The quality and quantity of all DNA extracts were evaluated (Buytaers et al., 2020) using the NanoDrop 2000 (Thermo Fisher Scientific), Qubit 3.0 fluorometer (Thermo Fisher Scientific) and 4200 TapeStation (Agilent). All DNA extracts were further processed using the Nextera XT library preparation kit (Illumina) before sequencing on the Illumina MiSeq, generating paired-end 250 bp reads with the reagent kit v3. The samples were sequenced in one run of eight libraries. The number of (paired-end) reads sequenced per metagenomics sample is presented in Table 4.1. Sequencing metrics were obtained using FastQC version 0.11.7 (Andrews, 2010).

Table 4.1: Quality metrics of the metagenomics sequencing and metagenomics assemblies

	Sauce	**Meal**
Sequencing metrics		
Total reads	2,653,700	4,857,796
Sequences flagged as poor quality	0	0
Sequence length	35-251	35-251
% GC	49	47
mean quality score	35.83	36.1
median quality score	30	31
Strain assembly metrics*		
# contigs	78	75
Largest contig	325,096	325,086
Total length	4,703,829	4,704,090
GC (%)	52.13	52.13
N50	106,626	128,74
Mean coverage	93.9	88.35
Median coverage	73.5	65.5

*Statistics based on contigs of size ≥500 bp.

4.2.4. Isolate data

Sequencing data from *S. enterica* subsp. *enterica* serovar Enteritidis isolates (see Table S1, available with the online version of this article) included data from five isolates of the hotel-school outbreak from food origin (the leftover meal and the three components of this meal that were all probably contaminated through spreading of the sauce between the compartments, and a chicken-based meal consumed on September 24th 2019 at the hotel school that was probably contaminated in the rubbish bin) and from five isolates from human origin linked to the hotel-school outbreak, obtained following conventional methods (EFSA Panel on Biology Hazards (BIOHAZ), 2014). These 10 isolates showed the same MLVA profile. As background for the phylogenetic analysis, data were also included from isolates linked to the still ongoing Polish outbreak (Pijnacker et al., 2019; ECDC and EFSA, 2020), presenting distinct MLVA profiles, i.e. seven Belgian isolates from food origin, five Belgian isolates from human origin and four isolates from public databases representing the different outbreak clusters defined by the Public Health England SNP pipeline described in an outbreak assessment from the European Centre for Disease Prevention and Control (ECDC) and EFSA (ECDC and EFSA, 2020), supplemented with ten isolates of human origin from Belgian sporadic cases from 2019, also presenting a different MLVA profile to the one of the hotel-school outbreak.

4.2.5. Data analysis

The metagenomics sequencing data were analysed through the workflow presented by Buytaers et al. (Buytaers et al., 2020): after trimming, a taxonomic classification of all reads to the genus level was performed using Kraken2 (Wood et al., 2019) (same databases as previously described (Buytaers et al., 2020)) in order to obtain an overview of the taxa present in the sample. The taxonomic classification results from Kraken2 (Wood et al., 2019) were verified using the online tools PathogenFinder (designed for isolate WGS) (Cosentino et al., 2013) using the model created for all bacteria, as well as CCMetagen (Marcelino et al., 2020a) used with the National Center for Biotechnology Information (NCBI) nucleotide database. Then, a strain-level read classification was performed using Sigma (Ahn et al., 2015) on a database of 787 complete genome assemblies of *Salmonella* (all serovars) from NCBI (list available upon request), using the default parameters as described by Saltykova et al. (Saltykova et al., 2020) to obtain the reads of the pathogenic strain, as *Salmonella* was the only pathogen detected after analysis of the taxonomic classification results. These reads as well as the sequencing reads from all isolates were assembled using SPAdes 3.13.0 (Bankevich et al., 2012). Quality metrics from the assemblies (Table 4.1) were obtained using quast version 5.0.2 (Gurevich et al., 2013). All assemblies from isolates and metagenomics samples were then typed (serovar prediction) using the online *Salmonella In Silico* Typing Resource (SISTR) (Yoshida et al., 2016) and the presence of AMR genes was detected using blast 2.6.0 on the ResFinder database (Zankari et al., 2012), with a minimum identity threshold of 90 % and a minimum length of 60 % for metagenomics assemblies, and 90 % minimum identity and minimum length for isolate assemblies (Bogaerts et al., 2019). The parameters were lowered for the metagenomics assemblies compared to the parameters (90 % gene coverage and 90 % nucleotide identity) chosen for the study of isolates, considering the lower depth obtained with metagenomics sequencing. For phylogenetic analysis, SNP calling was carried out on the classified (unassembled) reads as previously described (Saltykova et al., 2020), with *S. enterica* subsp. *enterica* serovar Enteritidis strain EC20120200 (*Enterobacteria*) as a reference

(GenBank accession no. CP007434.2). Maximum-likelihood substitution model selection and phylogenetic tree inference were done with mega (Kumar et al., 2018), using the NNI (nearest-neighbour-interchange) heuristic method, keeping all informative sites and using a bootstrap method with 100 replicates. The model selected to build the phylogenetic tree was that of Tamura and Nei (Tamura and Nei, 1993). iTOL (Letunic and Bork, 2016) was used for the representation of the tree, with the percentage of the reference genome covered annotated on each branch.

4.3. Results

4.3.1. Taxonomic classification of the metagenomics samples

Two food samples (meal and sauce component) that could be related to the outbreak after a first screening (culture and qPCR) were tested using a shotgun metagenomics approach in parallel to the conventional outbreak investigation carried out at the NRL. After culture-based enrichment of the food matrices, the DNA was extracted and sequenced. The reads obtained were then taxonomically classified to determine the genera that were present in the food matrices.

Only bacteria could be detected in both samples (89 and 96 % of the sequenced reads for the meal and the sauce, respectively), although the meal consisted of fish, mashed potatoes and sauce, and the sauce was made with fresh eggs. This was expected as the latter species (fish, potato and chicken) are not represented in the taxonomic databases used and, therefore, should be part of the unclassified section of the reads (Fig. 4.1). The same bacterial genera were detected in both matrices albeit at different relative abundances, except for *Streptococcus*, which was only present in the meal sample. The consensus in detected bacterial genera was to be anticipated since the sauce was sampled from the meal. *Salmonella*, the genus implicated in the outbreak, was detected at a high percentage in both matrices (70 % in the sauce, 40 % in the meal). This is consistent with the qPCR detection of the *Salmonella*-specific *invA* and *rpoD* genes in the DNA extracts of both samples (Table S2). However, other detected genera like *Escherichia*, *Bacillus*, *Klebsiella* or *Streptococcus* may also represent pathogenic species. Therefore, in an attempt to use the taxonomic classification as an agnostic tool to identify the causative food-borne pathogen, two other data analysis tools were used to determine the presence of a pathogen in the sample (CCMetagen and PathogenFinder). CCMetagen and PathogenFinder identified *S. enterica* as the main or only pathogen in the two samples (the results are shown in Table S3) after analysis based on KMA sequence alignments on the NCBI nucleotide database (CCMetagen) or prediction of pathogenicity based on the detection of groups of genes associated with human pathogenic bacteria (PathogenFinder). The output of the three different tools used, based on different bioinformatics approaches, confirmed that *Salmonella* was considered as the only pathogen meriting further investigation in this study.

4.3.2. Salmonella strain inference from metagenomics samples and in silico typing

Obtaining strains from the metagenomics reads is necessary to mimic the recovery and characterization of an isolate with conventional methods. This was done for each metagenomic sample following a previously reported metagenomics strain-level analysis pipeline (Buytaers et al., 2020; Saltykova et al., 2020). After classification of the reads to a database of *Salmonella* genomes, 1843873 and 1618032 reads were classified as ASM303203v1 [*S. enterica* subsp. *enterica* serovar Enteritidis (*enterobacteria*), RefSeq accession no. GCF_003032035.1], respectively, for the meal and the sauce (Table S4). This represents 38 % of the total sequenced reads for the meal and 61 % of the reads for the sauce. Less than 7000 reads (<0.5 % of the total reads) were classified to other *Salmonella* genomes for both samples, indicating that most probably only one strain of this species was present in the sample and that the reads assigned to ASM303203v1 correspond to that strain.

Consecutively, a sequence-based characterization can be performed on the reads of each inferred strain, corresponding to the characterization of the isolate with conventional methods. The reads were assembled (Table 4.1) and then typed *in silico*. The results (Table S5) confirmed that the strains obtained are indeed *S. enterica* subsp. *enterica* serovar Enteritidis,

based on O- and H-type prediction (serogroup D1, H1 g, m, H2-), multilocus sequence typing (MLST) clustering (ST11) and matches of their closest public genome. When comparing to the *in silico* typing of sequenced isolates from food and human origin from the outbreak (Table S5), the results were identical except for the detection of all 330 whole-genome MLST alleles in the isolates and 329 identical alleles in the metagenomics-based strains (one allele present partially). Other isolates obtained from the NRC, the NRL and from another outbreak circulating in Europe (not related to the hotel-school outbreak) were typed with the same tool. These were also defined as *S. enterica* subsp. *enterica* serovar Enteritidis, but were related to other genomes from public databases (Table S5).

The presence of AMR genes was also investigated in the assembled contigs of the metagenomics-based strains (Table S6), to follow the analysis that is usually performed on isolates (using the technique of microdilutions in broth), but then at the genotype level. The locus *aac(6')-Iaa_1*, linked to resistance to aminoglycoside due to a chromosomally encoded aminoglycoside acetyltransferase, was detected in all strains from the hotel-school outbreak, including strains derived from metagenomics sequencing, as well as all non-outbreak-related strains included in this study with 96.35 % identity and 100 % coverage (Table S6). The prevalence of this gene in *S. enterica* WGS from NCBI is 29 % (Mcarthur et al., 2013). No other AMR genes were detected in any strain.

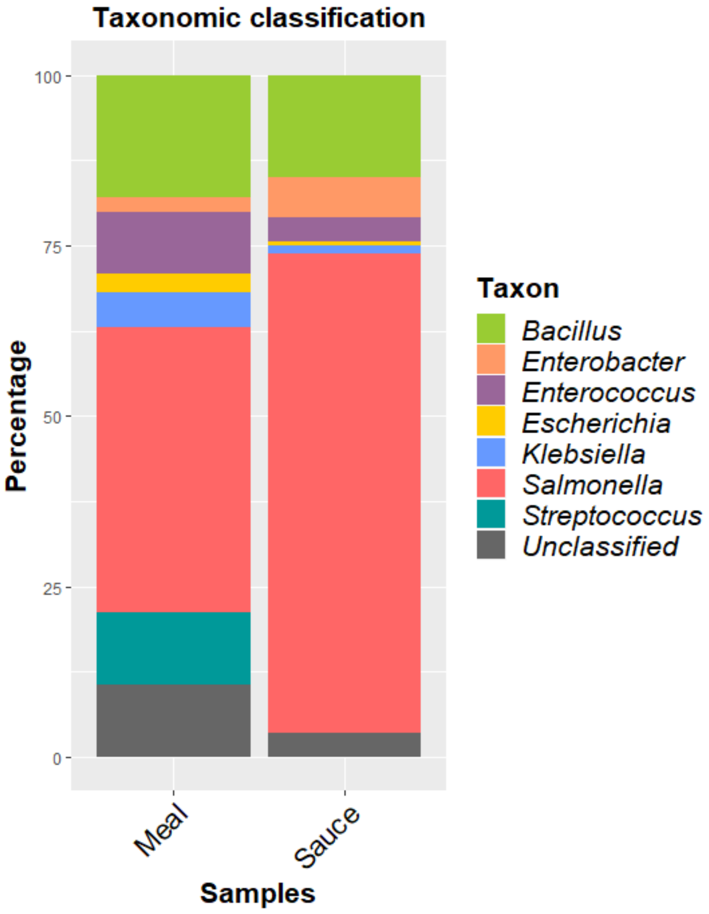


Figure 4.1: Percentages of reads classified to the genus level using a taxonomic classification tool (Kraken2) from metagenomics samples (full meal and sauce) with in-house databases of mammals, archaea, bacteria, fungi, human, protozoa and viruses. *Red represents the proportion of ‘ Salmonella ’ in the samples. The reads that could not be classified to the genus level for mammals, archaea, bacteria, fungi, human, protozoa or viruses are represented in grey.*

4.3.3. Metagenomics-based trace back investigation of the outbreak to its food source

Finally, in order to relate cases from food and human origins, the MLVA profiles can be compared with traditional methods, but EFSA now recommends WGS of *Salmonella* isolates and uses core-genome MLST in data sharing platforms such as EPIS. In our analysis, all isolates and metagenomics-derived strains were compared using SNP calling and reconstruction of a phylogenetic tree (Fig. 4.2). SNP calling offers the possibility of comparing the full genome and is considered more suited to use for metagenomics-derived strains (Saltykova et al., 2020). The cluster corresponding to the hotel-school outbreak (represented in blue in Fig. 4.2) includes the isolates from patients and suspicious food vehicles obtained by the NRC and NRL, as well as the two inferred strains obtained from direct sequencing of two food samples (suspect meal and sauce) using a shotgun metagenomics approach. The breadth of coverage of the reference genome for the two reconstructed strains from metagenomics samples is 97 and 85 % for the sauce and the meal, respectively. These values are in the same range as the values obtained for the isolates of the same outbreak. All strains of the hotel-school outbreak cluster, including the strains from the metagenomics samples, have 0 SNP differences per million genomic positions (Table S7). Other *S. enterica* subsp. *enterica* serovar Enteritidis circulating in Europe at the same time period, including isolates linked to an outbreak of Polish origin that started in 2016 but was still ongoing (shown in purple in Fig. 4.2), were included in the analysis, and could be separated both from the isolates and the metagenomics strains from the hotel-school outbreak.

4.3.4. Timing for a conventional and a metagenomics-based approach to resolve outbreak investigation to the food source

A schematic representation of the theoretical timeline of the conventional analysis conducted at the NRL on food samples, in parallel to the investigation on human samples conducted at the NRC, is presented in Fig. 4.3 (upper line). After receipt of the samples, the confirmation of the presence of *Salmonella* in the food is first conducted with qPCR on the food matrices, then normally isolates are obtained after approximatively 1 week (if isolates can be produced from the food samples), and characterized for serotype and MLVA profile. Once the MLVA profile is confirmed to be identical to the one detected in the patients' isolates, the DNA of the food isolates is extracted for WGS analysis. At the Belgian NRL, the serotyping and MLVA profile of the food isolates, if obtained, are currently prerequisites before sequencing, to prove that the strains have a high chance of being linked to the outbreak, as only outbreak cases are eligible for obtaining budget and priority for WGS. Notably, the isolates from human origin are most often already characterized at that stage as they are detected and isolated most often more easily and earlier in the investigation process. Together with library preparation, the sequencing takes approximately 4 days. The sequencing typically occurs 2 to 3 weeks after receipt of the samples depending on the isolation time, the

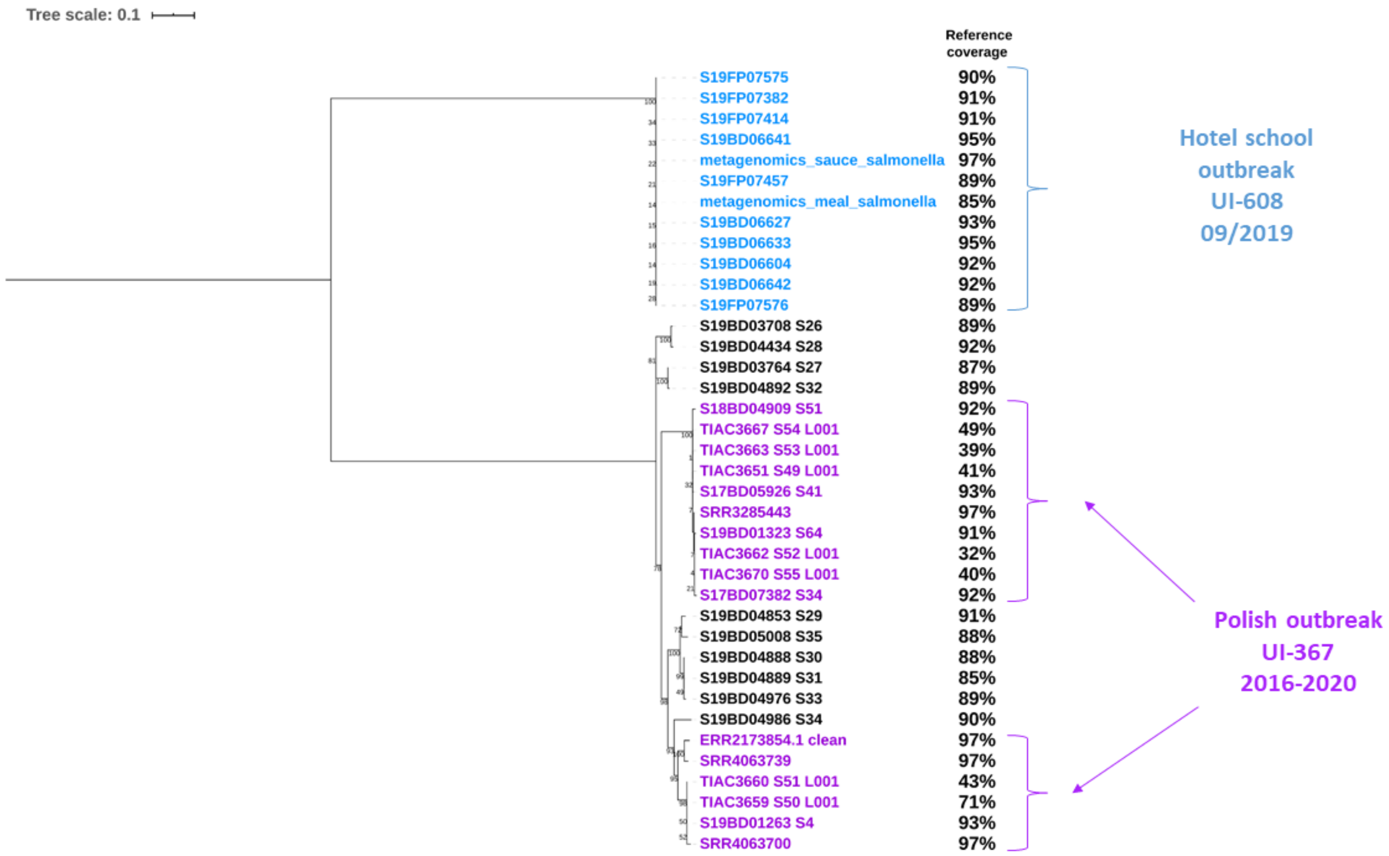


Figure 4.2: SNP-based phylogenetic tree representing the isolates and metagenomics-derived strains from food samples linked to the hotel-school outbreak (UI-608, in blue) in the global context of S. enterica subsp. enterica serovar Enteritidis circulating in Belgium and in Europe during the same time period. *Isolates linked to the Polish outbreak (UI-367) are indicated in purple, and isolates from sporadic cases in Belgium in 2019 in black. Percentage of the reference genome covered is presented on the side of each branch. Bar, nucleotide substitutions per 100 nucleotide sites. Node values represent bootstrap support values.*

time necessary to gather sufficient isolates to be cost-efficient for multiplexing in a single sequencing run, and to perform the sequencing run. Data analysis is then conducted, followed by sharing of the information, with national and international instances (in this case: RASFF 2019.3675 on October 16th 2019 and EPIS UI-608 updated on October 24th 2019 with the NGS data). In this outbreak, it allowed determination of the source of the contamination as an egg-producing farm in Spain and detection of 13 related human cases from France and 2 human cases in both the Netherlands and the UK (Sciensano, 2020). In the same time period, an outbreak was reported in the Netherlands involving eggs originating from Spain (RASFF 2019.3069, UI-601). However, the strains of *S .enterica* Enteritidis had distinct MLVA profiles, 2-11-7-3-2, 3-10-5-4-1, 2-10-7-3-2, 3-11-5-4-1, and 170 core-genome MLST allelic differences from our outbreak strain. The UK also reported an outbreak linked to eggs (RASFF 2019.1412, UI-602), but again no link with the Belgian outbreak strain was established. The WGS data of these strains were not publicly available and, therefore, could not be added to the phylogenetic analysis in this study.

This timeline was compared to that of a metagenomics-based analysis of the food samples. DNA from the meal and the sauce was extracted from a small fraction of the cultured food matrices for subsequent metagenomics analysis after suspicion of the contamination with qPCR (not necessary for a metagenomics-only workflow). From the time of the DNA extraction, depending on the availability of a sequencing instrument and the preparation of the libraries, the sequenced reads could be obtained in a minimum of 4 days (Fig. 4.3, lower line). Thereafter, a taxonomic classification was obtained in a few minutes and, after 1 day, a pathogenic strain was obtained and fully typed. In less than a week after receipt of the samples in the laboratory, the pathogen was fully described and related to other cases from the outbreak (from food and human origin) in a phylogenetic tree. This corresponds already to the mean time necessary to only obtain an isolate from food in routine analysis, if obtained, with no information about relatedness of the cases at that stage of the conventional analysis. Indeed, in the conventional analysis, obtaining a food isolate is a prerequisite for performing the molecular analysis, including WGS, to be able to determine relatedness.

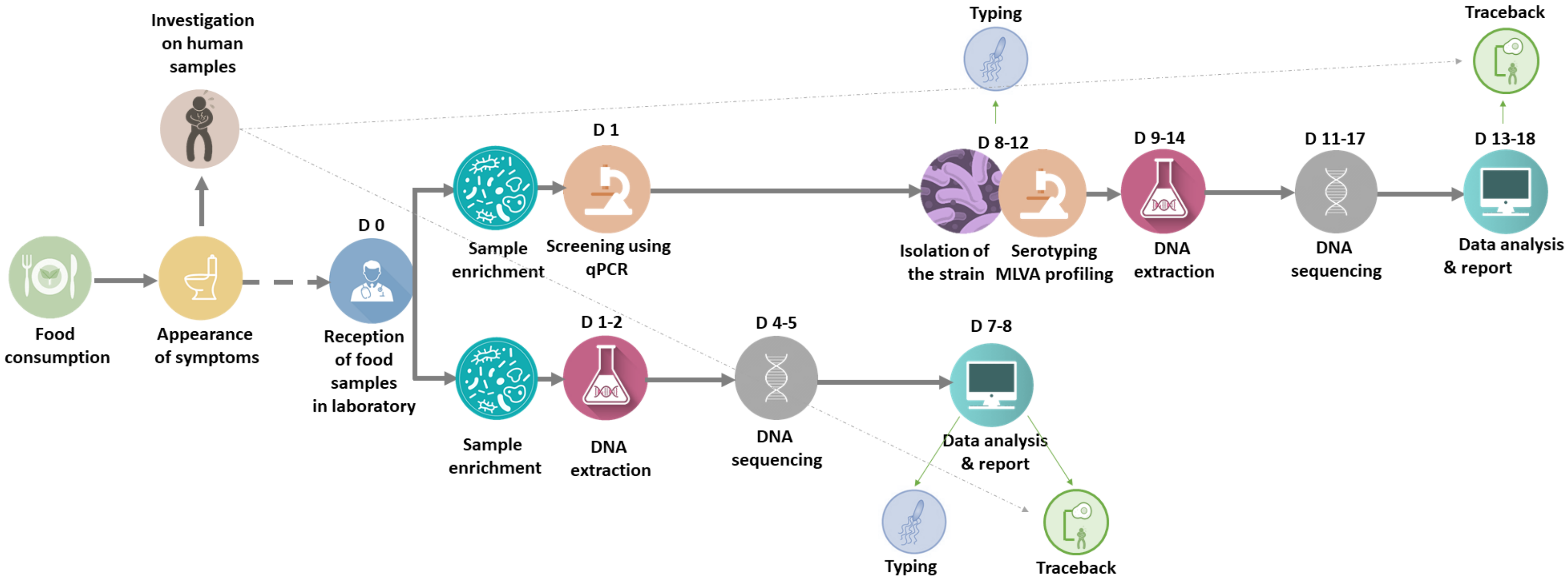


Figure 4.3: Comparison of theoretical processing time for the conventional approach (upper level) and the shotgun metagenomics approach (lower level) for Salmonella -contaminated food samples from receipt of the samples to strain typing and trace back between human and food strains. *A range of days (D x–y) accounts for a range of duration of some laboratory analyses, which can vary due to the presence of technicians during weekends, success in the isolation process or cost-effectiveness (start of the sequencing run with sufficient samples)*

4.4. Discussion

We deliver in the present study a proof of concept for the shotgun metagenomics approach on food samples previously developed on food samples artificially spiked with STEC (Shiga toxin-producing *E. coli*) (Buytaers et al., 2020) to resolve a *Salmonella* outbreak in Belgium up to the food source. We described the analysis of an outbreak that affected over 200 students and teachers at a hotel school in Belgium, using a strain-level shotgun metagenomics-based approach in parallel to the investigation based on WGS of isolates performed by the NRL and NRC. Two suspect samples of leftovers of the meal and the tartar sauce included in this dish were analysed with a shotgun metagenomics workflow, in a relatively very short time frame, and the pathogenic strain was inferred from the sequenced metagenomics reads and characterized as a *S. enterica* subsp. *enterica* serovar Enteritidis that was related with 0 SNP differences to the isolates of human origin from the same outbreak. Therefore, the outbreak could be resolved, i.e. source attribution, using metagenomics data for the food samples. As this was a proof of concept, isolates were also obtained and characterized from the food samples through conventional analysis, and were also related to the metagenomics strains with 0 SNP differences, as a validation of the obtained results. Moreover, the outbreak cluster was placed in a global perspective of the situation of salmonelloses in Belgium and Europe using a phylogenetic tree including other strains circulating at the same time period.

The timing of an outbreak investigation is a critical factor to limit the propagation of the contamination. Shotgun metagenomics is an alternative to the conventional approaches circumventing the need for isolation, which is time-consuming and most importantly not always achievable in routine analysis. This study showed the potential of metagenomics to be used during outbreak investigations on food samples for obtaining the same level of information as from food isolates, in a time frame reduced by over 1 week. Moreover, this constitutes a pathogen-agnostic approach dependent on a non-selective enrichment, which allows the detection of the pathogenic strains (here *Salmonella*) and the characterization of this contaminant without prior knowledge on the species or the number of different species and/or strains present in the sample (Sciensano, 2020), in contrast to conventional methods where the assumption of the species to test for is based on the symptoms of the patients. Therefore, this metagenomics approach is also advantageous in case of a limited quantity of food leftovers, because no choice for best fit symptoms-pathogen should be made as for conventional methods. Hence, this approach can potentially increase the range of pathogens detected in a mixed sample, and help reduce even more the economic burden of such food-borne pathogens, as was already stated for WGS of isolates (Jain et al., 2019). Our approach still relies on the isolation of the pathogen from the human samples and is not a stand-alone metagenomics approach. As the bacterial load is generally higher in human samples, isolation is not reported as a challenge in these matrices. Moreover, the isolation in the human samples is often not a limiting factor for the timing of food-borne outbreak investigation, as these samples are often obtained before the food samples in the case of outbreaks. Nevertheless, metagenomics studies of stool samples, included during outbreaks, have been published previously (Loman et al., 2013; Quick et al., 2015; Huang et al., 2017), and such an approach could be performed in parallel to the one we present, in the corresponding institution (NRC). However, this would represent a higher cost and the sequencing of human DNA might lead to ethical and privacy issues, in particular in Europe.

At a national scale, the typing data of food and human isolates are shared between the NRL and NRC, and matches are reported at the European level, i.e. EFSA and ECDC (Sciensano, 2020). No shared database is publicly available at the moment and access to this data or the samples must go through contact between both national entities. Communication concerning human health at the international level for outbreaks in Europe is done through the use of a communication platform and data sharing between public-health experts, by 'Urgent Inquiries' at the EPIS platform. For food safety, communications are done by the competent authorities through the RASFF system. These tools were used in the hotel-school outbreak investigation and helped to trace back and link the outbreak to eggs originating from Spain and other human cases in France, the UK and the Netherlands ECDC (Sciensano, 2020). However, for confidentiality reasons, these data were not made publicly available and, therefore, could not be included in our presented phylogenetic tree. Our study highlights that access to scientific data, including both raw WGS data and processed results, from public-health and food-safety authorities at both the national and international level will help to strengthen analyses on international outbreaks such as the one presented in this study, and consequently should be considered in the line of data sharing systems that have proven their efficiency.

The shotgun metagenomics approach has proven its potential for outbreak investigation through studies like this one, yet additional research could help with the actual further implementation of this method in routine settings. First, the culture of the food matrix as currently specified in the ISO (International Organization for Standardization) method could be adapted to suit a larger number of species concurrently for pathogen-agnostic metagenomics studies. Second, the optimal quality-control metrics for metagenomic sequencing have not yet been established, in contrast to ongoing efforts for WGS of isolates (e.g. ISO/DIS 23418 (ISO: International Organization for standardization)). In the current analysis, eight metagenomic food samples (six were not related to this study) were multiplexed in a single MiSeq run, with a relatively high cost per sample as a result. This allowed achievement of a sequencing depth of >85× for the single detected *Salmonella* strain for both metagenomic samples, which is comparable to values typically achieved for isolates and is more than sufficient for the reconstruction of the pathogen's genome. This indicates that, in the future, sequencing of a higher number of samples simultaneously can be attempted, lowering the cost. The observed coverage is, however, much higher than in our previous work, where multiplexing of 12 minced meat samples resulted in sequencing depths between 0.9× and 10× for detected *E. coli* strain(s) (Buytaers et al., 2020). Leonard et al. (Leonard et al., 2015, 2016) reported that multiplexing of 12 enriched spinach samples yielded coverages between 5× and 145× for an *E. coli* reference genome, with 4 samples having coverages less than 30×. Therefore, the minimal required sequencing depth will likely differ for each sample type, and will depend on biological factors such as the initial load of contamination or the efficiency of the enrichment procedure, and the expected number of bacterial strains. Generally, we have observed that coverages of over 5–10× can be sufficient for detection of virulence genes and phylogenetic placement of bacterial strains in case reference-based assembly is used (Saltykova et al., 2020). However, there is a need to precisely establish the reliability of the strain characterization and subtyping results obtained using data of different sequencing depth. Third, user-friendly pipelines need to be developed to be used directly in the laboratory by non-expert bioinformaticians. Moreover, bioinformatics taxonomic identification tools should be further tested and improved, so that different tools, each with their advantages and limitations, provide the same results, and to avoid misclassifications (Marcelino et al., 2020b). However, the focus of this study was not to present a benchmarking of bioinformatics tools for strain-level shotgun metagenomics, but rather a proof of concept based on previously developed bioinformatics methodologies (Buytaers et al., 2020; Saltykova et al., 2020). Other approaches and tools might still improve the results (accuracy, speed of analysis) and could be evaluated in further studies (Seeman, 2015; Minh et al., 2020). This confirms the need for studies such as this one to produce data to make benchmarking analyses possible or help in the design of new tools. Another perspective for the implementation of this method in routine analysis is the reduction of the analysis cost. As elaborated above, shotgun metagenomics analyses imply runs with a very limited number of samples on Illumina sequencers in order to maximize the sequencing depth. Other sequencing devices as manufactured, for instance, by Oxford Nanopore Technologies offer real-time long-read sequencing of one sample at a time, at a low price if using the Flongle flow cell. Such fast sequencing could also further reduce the turnaround time of a metagenomics-based outbreak investigation (Juul et al., 2015). However, its applicability for strain-level characterization in complex samples remains to be demonstrated.

In 2019, the EFSA published an opinion on the use of metagenomics for outbreak investigation (EFSA, 2019b), describing the possibilities offered by an isolation-free method. However, at that time, metagenomics had not yet been used to resolve a food-borne outbreak investigation to its food source and was considered as experimental. Moreover, it was considered technically challenging to obtain a draft genome of the pathogenic strain in order to assign particular genetic determinants to the causative agent. This study has shown that a *Salmonella* outbreak caused by a complex food matrix could be resolved to strain resolution using shotgun metagenomics, in a shorter time frame than needed for isolation of the strain, paving the way for future studies to use this method outside the experimental scope and to support the EFSA opinion.

Acknowledgements

We would like to thank all collaborators from the agency Zorg en Gezondheid, the Federal Agency for the Security of the Food Chain and the Sciensano epidemiology department who participated in this outbreak investigation. We are grateful to the technicians from the NRL and NRC of *Salmonella* (Sciensano) for their work on the bacterial isolates and the food enrichments, as well as Stefan Hoffman, Maud Delvoye and Els Vandermassen from Transversal Activities in Applied Genomics (Sciensano) for their involvement in the sequencing runs. Finally, we acknowledge Mathieu Gand for sharing his expertise about *Salmonella* and serotyping.

Supplementary data

The NCBI accession numbers for the new sequence data presented in this paper are SAMN15963373–SAMN15963404, SAMN15957185 and SAMN15957186 (see Table S1).

All supporting data, code and protocols have been provided within the article or through supplementary data files. Seven supplementary tables are available with the online version of this article.

CHAPTER 5
Towards Real-Time and Affordable Strain-Level Metagenomics-Based Foodborne Outbreak Investigations Using Oxford Nanopore Sequencing Technologies

This chapter was adapted from the previously published manuscript:
Buytaers, Florence E.*, Assia Saltykova*, Sarah Denayer, Bavo Verhaegen, Kevin Vanneste, Nancy H. C. Roosens, Denis Piérard, Kathleen Marchal, and Sigrid C. J. De Keersmaecker. 2021. "Towards Real-Time and Affordable Strain-Level Metagenomics-Based Foodborne Outbreak Investigations Using Oxford Nanopore Sequencing Technologies." Frontiers in Microbiology 12 (November): 1–13. https://doi.org/10.3389/fmicb.2021.738284. (* equal contribution)

Authors' contributions:

F. E. Buytaers, A. Saltykova, and S. De Keersmaecker conceptualized the study, conducted the formal analysis, conducted the investigation, were responsible for the methodology, and wrote the original draft. F. E. Buytaers designed and performed the wet lab (spiking), performed the strain-level data analysis and the *in silico* DNA walking. A. Saltykova designed and implemented the SNP data analysis workflow. F. E. Buytaers, S. Denayer, B. Verhaegen, and D. Piérard curated the data. N. H. C. Roosens and S. De Keersmaecker were involved in the funding acquisition and were involved in project administration. F. E. Buytaers, S. Denayer, B. Verhaegen, K. Vanneste, N. H. C. Roosens, and D. Piérard provided the resources. F. E. Buytaers, A. Saltykova, and K. Vanneste were responsible for the software. K. Marchal and S. De Keersmaecker supervised the study. S. Denayer, N. H. C. Roosens, D. Piérard, and S. De Keersmaecker performed the validation. F. E. Buytaers was responsible for the visualization. All authors reviewed and edited the manuscript, contributed to the article, and approved the submitted version.

Abstract:

The current routine laboratory practices to investigate food samples in case of foodborne outbreaks still rely on attempts to isolate the pathogen in order to characterize it. We present in this study a proof of concept using Shiga toxin-producing *Escherichia coli* spiked food samples for a strain-level metagenomics foodborne outbreak investigation method using the MinION and Flongle flow cells from Oxford Nanopore Technologies, and we compared this to Illumina short-read-based metagenomics. After 12 h of MinION sequencing, strain-level characterization could be achieved, linking the food containing a pathogen to the related human isolate of the affected patient, by means of a single-nucleotide polymorphism (SNP)-based phylogeny. The inferred strain harbored the same virulence genes as the spiked isolate and could be serotyped. This was achieved by applying a bioinformatics method on the long reads using reference-based classification. The same result could be obtained after 24-h sequencing on the more recent lower output Flongle flow cell, on an extract treated with eukaryotic host DNA removal. Moreover, an alternative approach based on *in silico* DNA walking allowed to obtain rapid confirmation of the presence of a putative pathogen in the food sample. The DNA fragment harboring characteristic virulence genes could be matched to the *E. coli* genus after sequencing only 1 h with the MinION, 1 h with the Flongle if using a host DNA removal extraction, or 5 h with the Flongle with a classical DNA extraction. This paves the way towards the use of metagenomics as a rapid, simple, one-step method for foodborne pathogen detection and for fast outbreak investigation that can be implemented in routine laboratories on samples prepared with the current standard practices.

5.1. Introduction

Foodborne diseases represent a major burden worldwide (WHO, 2015). Foodborne pathogens can cause large outbreaks affecting multiple people sometimes in different regions. In case of an outbreak, the common practice of public health institutions is to investigate human cases and try to relate them to the contaminated food, in order to remove it from the food chain and prevent further contaminations. This process is called source attribution (EFSA, 2019a). This investigation consists of a microbiological and epidemiological part. In many countries, a surveillance system is also in place, screening the food chain in order to remove contaminated foodstuffs before they reach the consumer. In that case, microbial risk assessment and hazard identification are conducted, and the pathogen does not need to be linked to patient's data, but its characteristics could be added to a database in order to conduct retrospective studies and link related cases or serve as background to detect clusters and thus putative outbreaks outbreaks (ECDC and EFSA, 2019).

In both circumstances (i.e., surveillance or the microbiological part of the outbreak investigation), conventional microbiology methods based on sequential culture steps have been the standard for many years to obtain information on the bacterial contaminant(s) present in food. However, this depends on a series of steps that should be conducted on the samples, therefore requiring larger quantities of the sample that is not always easy to obtain, and most importantly, it requires obtaining an isolate, which is often time-consuming and not always successful. The heterogeneous contamination of food products, the complexity of the matrix, and the difficulty to culture certain organisms might not allow to detect a pathogen at levels as low as the infectious dose reported for human (Food and Authority, 2018). When an isolate is obtained, it is characterized with several (real-time) polymerase chain reactions [(q)PCRs] to detect pathogenicity markers and/or multiple locus sequencing typing (MLST), pulsed-field gel electrophoresis (PFGE), multiple locus variable-number tandem repeat analysis (MLVA), or other typing methods to relate cases of an outbreak, depending on the pathogen. This workflow does not always offer optimal resolution to discriminate the pathogenic agents at a desired level (Nouws et al., 2020a) and requires sequential tests to be conducted in the laboratory (Nouws et al., 2020a), which adds to the total cost and turnaround time of the analysis.

As an alternative, whole-genome sequencing (WGS) offers the ultimate resolution to the single-nucleotide polymorphism (SNP) level of the bacterial genome, allowing the simultaneous detection of all genes present in the bacterium as well as relatedness inference with phylogenetics (Sandora et al., 2014; Bogaerts et al., 2019), and has been recommended by the European Food Safety Authority (EFSA) for use on a list of pathogens in European laboratories (EFSA, 2014b). However, circumventing the need for isolation can accelerate the collection of results even more, as well as allow the resolution of cases for which no isolate could be obtained following the detection protocol. Strain-level shotgun metagenomics approaches offer the possibility to obtain the same resolution as WGS, without the need for isolation (Forbes et al., 2017). A recent publication of the EFSA highlighted the need for demonstrating the ability of metagenomics to be used as a new alternative for risk assessment, source attribution, and outbreak investigation (EFSA, 2019b).

In our previous work, we have presented a metagenomics approach to obtain the same level of precision as the conventional bacterial detection methods and isolate's WGS, through direct sequencing of all DNA in the sample after enrichment in a non-selective medium following the ISO standard ISO 13136:2012 (ISO: International Organization for standardization, 2012; Buytaers et al., 2020, 2021c). After short-read sequencing of 12 DNA extracts with or without removal of host DNA in a 48-h Illumina MiSeq run, we were able to link the pathogenic strains derived from metagenomics sequencing of samples containing multiple strains of the same species (*Escherichia coli*) to human isolates from the same outbreak (Buytaers et al., 2020). This was possible using a bioinformatics workflow classifying short reads to a reference genome database (Saltykova et al., 2020). Although Illumina is a widely used sequencing technology generating short reads with high accuracy, it still comes at a high cost for metagenomics, impeding a real implementation in routine. Moreover, the rather long library preparation time for multiple samples that have to be multiplexed to make the run cost-effective, as well as the 48-h sequencing run time, is not ideal for a fast response in case of an ongoing outbreak. Real-time long-read sequencing is now offered by Oxford Nanopore Technologies (ONT) with faster library preparation protocols coupled with the flexibility to cost-efficiently sequence one sample at a time on the flow cells. This could speed up the analysis of samples in an outbreak investigation and help to decrease the cost, which remains important, of metagenomics if using more cost-effective consumables for lower amounts of samples such as the MinION flow cell or the new lower output Flongle flow cell. Furthermore, long-read sequencing offers the possibility to investigate larger genome fragments without the possible bias of short-read metagenomics assembly, which could offer an added value in the context of metagenomics-based outbreak investigation.

Sequencing using Oxford Nanopore Technologies has been previously validated for the characterization of foodborne pathogenic isolates, even during the course of an outbreak (Loman et al., 2015; Quick et al., 2015; Greig et al., 2019), and has since then been tested in some metagenomics studies for pathogen identification by species and gene detection in the mixed reads (Schmidt et al., 2017; Charalampous et al., 2019). It was shown to allow attribution of potentially pathogenic taxa to the corresponding antimicrobial resistance genes they harbored by gene walkout (Leggett et al., 2020). However, strain-level characterization is necessary for the precise resolution of an outbreak, which remains a challenge for ONT metagenomics data partly due to the higher error rate of the technology (Forbes et al., 2018; Gardy and Loman, 2018). In a previous study, Hyeon et al. (2018) used an enriched food sample that was artificially contaminated with *Salmonella*, treated with immunomagnetic separation to concentrate the target bacteria, and whole-genome amplification before it was sequenced using the MinION technology. They obtained 65 and 70 SNP difference to the WGS isolate reference of the spiked bacterium after 1.5 and 48.5 h of sequencing, respectively (Hyeon et al., 2018). A similar quasimetagenomics method was used to target Shiga toxin-producing *E. coli* (STEC) and *Salmonella* in contaminated flour samples (Forghani et al., 2020).

The method proved successful to cluster (without specifying the SNP differences) the metagenomics-obtained strain to the spiked isolate, for multiple single-spiked strains of each pathogen and also on samples co-spiked with one strain of each of the two pathogen species. However, this approach is still rather new, and new proofs of concept are necessary to demonstrate that it can be effectively used, possibly with a lower amount of SNP differences, for more reliable cluster definition in daily outbreak investigation. Indeed, it has not yet been tested with a non-selective enrichment method, a procedure closer to the ones currently followed by the reference laboratories (UE, 2005). Moreover, it has not yet shown its efficiency not only in samples possibly presenting multiple strains of the same species but also to cluster the metagenomics-derived strain to related human cases from the same foodborne outbreak. Finally, sequencing not only on the lower cost but also lower output, Flongle flow cell device still remains to be evaluated for such an application.

We present in this study a proof of concept of shotgun metagenomics outbreak investigation performed after ONT sequencing, combined with a new bioinformatics workflow adapted to long reads, to obtain the characterization of the foodborne pathogen at strain level in samples with various strains of the same pathogen (STEC). The spiked food samples were previously sequenced on Illumina and reported in former studies (Buytaers et al., 2020; Saltykova et al., 2020). A comparison between the results obtained with the two sequencing technologies was made. Moreover, a new approach, *in silico* DNA walking, offering the screening of food samples for pathogens at low cost based on long reads after Flongle sequencing, was evaluated after DNA extraction with or without host DNA removal. Finally, a strategy to integrate metagenomics in the current screening and pathogen characterization at the routine laboratories was proposed based on the results obtained after Flongle, MinION, and Illumina sequencing and their respective cost-effectiveness and execution time.

5.2. Materials and methods

5.2.1. Selection of the samples

Minced beef meat harboring a natural population of commensal *E. coli* bacteria and artificially contaminated with a low infection dose of STEC from a previous study (Buytaers et al., 2020) was used to evaluate the performance of MinION and Flongle sequencing compared to Illumina MiSeq sequencing on the same sample. Briefly, 25 g of the food matrix spiked with 5 colony-forming units (CFU) of STEC was enriched in buffered peptone water for 24 h at 37°C, following the culture described in ISO 13136:2012 for STEC detection in food (ISO: International Organization for standardization, 2012) in order to be representative of the procedures followed by the reference laboratories and therefore the samples they could get to analyze. One milliliter of the mix was used for DNA extraction using the NucleoSpin Food kit (Macherey-Nagel, Düren, Germany) or HostZERO Microbial DNA kit (Zymo Research, Irvine, CA, United States). The latter is advertised as able to remove host DNA. The strain that was chosen to artificially contaminate the food matrix was a STEC O157:H7 eae+, stx1+, stx2+, isolated during an outbreak in Limburg, Belgium, in 2012 (Braeye et al., 2014), and previously characterized through WGS (Nouws et al., 2020b). A negative control, a blank of the enriched food matrix, was previously sequenced on Illumina MiSeq and characterized to pinpoint the presence of commensal *E. coli* bacteria and the absence of STEC virulence genes in the meat prior to spiking (Buytaers et al., 2020; Saltykova et al., 2020).

5.2.2. Oxford Nanopore MinION Sequencing

The DNA library was prepared with the Genomic DNA by Ligation protocol (SQK-LSK109; Oxford Nanopore Technologies, Oxford, United Kingdom) on the DNA extracted with the NucleoSpin kit. It was performed according to the recommendations for MinION sequencing on a MinION flow cell (R9.4.1). The prepared library was then loaded on a primed flow cell (R9.4.1), and a 48-h sequencing run was started, generating 1.2 million reads with a median length of 1,991 bp. The resulting fast5 files obtained at various sequencing time checkpoints were basecalled using Guppy version 4.2.3 (Oxford Nanopore Technologies).

5.2.3. Oxford Nanopore Flongle Sequencing

Two DNA libraries were prepared, respectively, for the DNA extracted with the NucleoSpin and the HostZERO kits with the Genomic DNA by Ligation protocol (SQK-LSK109; Oxford Nanopore Technologies, Oxford, United Kingdom), following recommendations for Flongle sequencing. Each library was then loaded separately on a primed Flongle flow cell (R9.4.1), and a 24-h sequencing run was started, generating 244,019 and 187,966 reads with a median length of 686 and 3,393 bp, respectively, for the NucleoSpin and HostZERO DNA extracts. The basecalling was performed at various sequencing time checkpoints as in the "Oxford Nanopore MinION Sequencing" section.

5.2.4. Long-Read Strain-Level Metagenomics Data Analysis

First, a taxonomic classification with Kraken2 (Wood et al., 2019), using the same databases (in-house database of mammals, archaea, bacteria, fungi, human, protozoa, and viruses) as used for the Illumina analysis of the same samples (Buytaers et al., 2020), was performed on the basecalled reads of MinION and Flongle sequencing, including after specific time check-points. Graphs were created on the classification results using ggplot2 in R.

Second, the presence of virulence genes in the sequenced reads and the genomic context (taxon) of the same sequencing fragment were determined using an *in silico* DNA walking method, previously described for the detection of genetically modified microorganisms using a metagenomics approach (Buytaers et al., 2021a). Briefly, a Basic Local Alignment Search Tool (BLAST) analysis was performed on all reads using BLASTn version 2.7.1 with default parameters (Camacho et al., 2009) to the databases VirulenceFinder *E. coli* (Joensen et al., 2014) and nucleotide from NCBI (Bethesda (MD): National Library of Medicine (US), 1988). The hit to the NCBI database of each fragment presenting a virulence gene was used to obtain the genomic origin of the read harboring the virulence factors. The results were finally filtered to retain only the results for the virulence genes *stx1, stx2, eae*, and *ehxA*. For the goal of this study, focusing on a fast response to the detection of a foodborne pathogen, we presented the results obtained in the shortest timeframe necessary to obtain at least one read confirming the presence of a STEC in the sample. The results were visualized using sunburst charts.

Finally, the *E. coli* strains were inferred using Metamaps v 0.1 (Dilthey et al., 2019). Thereby, ONT reads from MinION and Flongle sequencing, including after specific time checkpoints, were classified against a database containing 2,831 reference sequences corresponding to the 976 complete *E. coli* genomes and complete 1,885 *E. coli* plasmids available from RefSeq on August 11, 2019 (O'Leary et al., 2016). Reads assigned to sub-species-level taxa were extracted.

A gene detection was conducted on the clustered reads of the inferred strains using BLAST version 2.7.1 (Camacho et al., 2009) on the VirulenceFinder *E. coli* database (Joensen et al., 2014) and SerotypeFinder O type and H type (Joensen et al., 2015) with default parameters. The strains containing stx genes were considered as STEC strains.

For the phylogenetic analysis, extracted reads of the STEC strain sequenced with ONT devices were mapped to a common STEC reference genome (BA000007.3) using bwa mem v 0.7.17 with the ont2d parameter set. Illumina sequences were previously analyzed through a similar workflow (Buytaers et al., 2020; Saltykova et al., 2020). Bcftools v 1.9 was used for the initial identification of potential SNPs as positions at which at least five reads contained an alternative allele, followed by filtering whereby positions with a minimal depth of 10 reads, a minimal allele frequency of 0.85, and a minimal mapping quality of 50 were retained (Supplementary Material 1). Genomic positions that did not meet the minimal sequencing depth and the minimal mapping quality criteria and potential SNPs that did not meet the minimal allele frequency were masked in the consensus sequence. Maximum likelihood substitution model selection and phylogenetic tree inference were performed using MEGA (Kumar et al., 2018), applying the nearest-neighbor-interchange (NNI) heuristic method, keeping all informative sites and using the bootstrap method with 100 replicates as a phylogeny test. The model selected was the Kimura two-parameter model with uniform rates among sites. Strains inferred from the Illumina sequencing of the same metagenomics samples [NucleoSpin extract and HostZERO extract (Buytaers et al., 2020)] and isolates from human (TIAC 1165 and TIAC 1169) and food (TIAC 1151 and TIAC 1152) originating from the same outbreak (Braeye et al., 2014), as well as some sporadic cases from the same serotype O157:H7 (TIAC 1638 and TIAC 1153), were used as background for the phylogenetic tree construction. All isolates were sequenced for a previous study (Nouws et al., 2020b). All workflows of command lines used for bioinformatics analyses in this work are presented in Supplementary Material 2.

5.3. Results

5.3.1. Long-Read Sequencing on a MinION Flow Cell for Strain-Level Metagenomics Outbreak Investigation

DNA extracted from beef meat spiked with STEC at the lowest infection dose was sequenced on a MinION flow cell. A data analysis workflow was developed in order to produce similar results as those generated with Illumina sequencing (Buytaers et al., 2020) and WGS of isolates, i.e., obtaining and characterizing the reads corresponding to the pathogenic strain in the sample and performing SNP-level phylogeny with this strain.

Beef ("Bos," blue) was the main species detected in the sample after both sequencing runs (Figure 5.1). This was to be expected as the sample consisted of beef meat. *Ovis* (a genus that includes sheep, olive green) was classified for a small part (2%) of the reads after MinION sequencing. The bacterial genera detected were identical between the two sequencing technologies. *Escherichia*, the pathogen not only artificially spiked in the sample but also endogenously present in the beef before spiking [Blank_Illumina (Buytaers et al., 2020)], was identified for 8 and 6% of the reads after Illumina and MinION sequencing, respectively. All species were detected after 30 min of sequencing on the MinION.

5.3.1.1. Confirmation of the Presence of a Pathogen in the Sequenced Metagenomics Sample Using in silico DNA Walking on Long Reads

In order to indicate the presence of a pathogen in the sample after Illumina sequencing, a virulence gene detection was conducted on all reads (Buytaers et al., 2020). However, with that information, the virulence gene cannot be linked to the pathogen's genome, which would be proof of the presence of the pathogen in the sample. Long-read sequencing offers the possibility to investigate the DNA fragment on which a virulence gene is detected in order to attribute it to a taxon (genomic context). This analysis is also known as *in silico* DNA walking.

As the sample was artificially spiked by a known STEC isolate, our approach was targeted at this pathogen specifically. Therefore, *in silico* DNA walking was applied to all long-read sequences with BLAST on the databases of *E. coli* virulence genes and nucleotides from NCBI, to determine if the *Escherichia*-related virulence genes, in particular *stx* genes defining an *E. coli* as a STEC pathogen, were found on *Escherichia* genome sequences, proving the presence of a pathogenic strain in the sample. This approach was tested as a fast alternative to obtain minimal characterization information on the pathogen in the sample before the inference of the strains from the metagenomics reads.

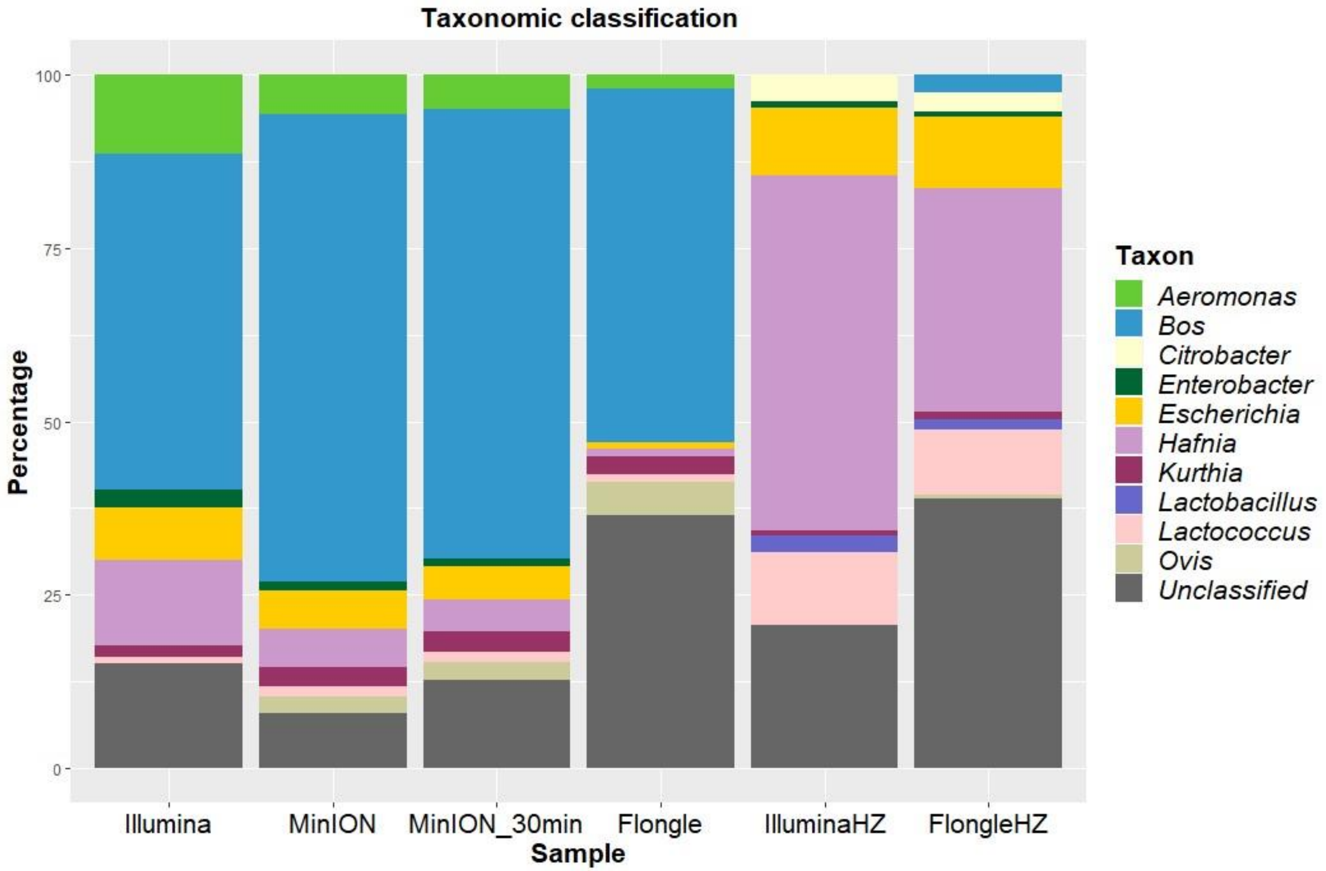


Figure 5.1: Percentages of reads classified to the genus level using Kraken2 (taxonomic classification tool) from blank and spiked beef samples extracted with two DNA extraction kits (one involving host removal, HZ) and sequenced on Illumina (MiSeq), MinION, or Flongle, with in-house databases of mammals, archaea, bacteria, fungi, human, protozoa, and viruses. *The data for Illumina sequencing (*) was published in Buytaers et al. (2020). Light blue represents the proportion of "Bos" corresponding to beef reads. Yellow indicates the presence of "Escherichia" in the sample. The reads that could not be classified to the genus level for mammals, archaea, bacteria, fungi, human, protozoa, or viruses are represented in gray.*

The results, presented in Figure 5.2, show that the virulence genes characteristic of the spiked STEC pathogen (*stx, eae*, and *ehxA*) could be linked to *Escherichia* fragments after already 1 h of MinION sequencing. This demonstrated that an *Escherichia* strain carried these genes, therefore indicating that STEC DNA was present in the samples. Moreover, as the enriched blank meat was previously sequenced and characterized (Buytaers et al., 2020), we can rule out the presence of STEC, *E. coli* virulence genes, or stx phages in the meat prior to the artificial contamination.

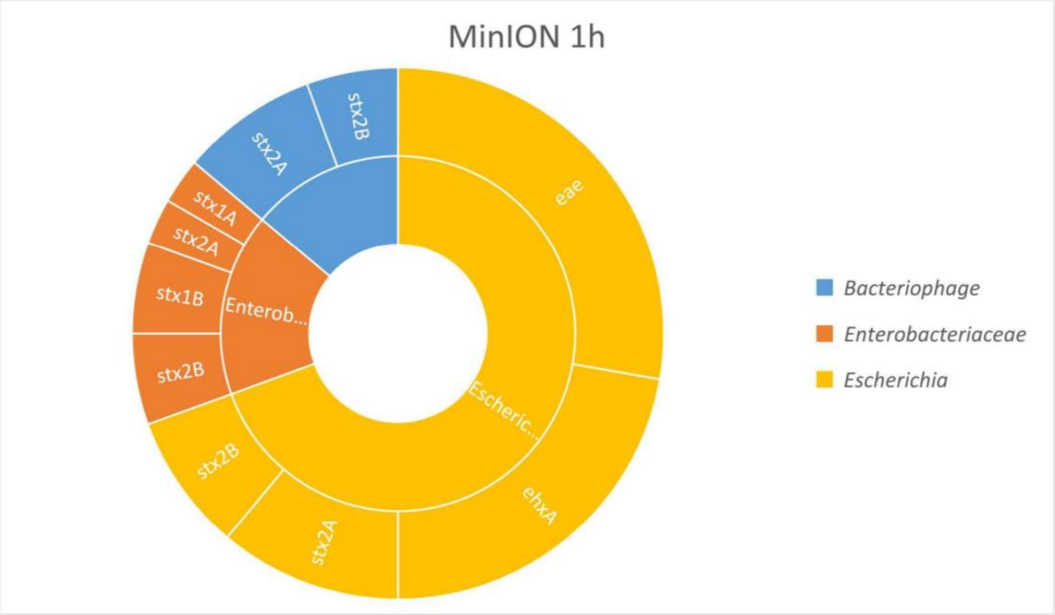


Figure 5.2: In silico DNA walking results, presenting the genera in the inner circle (following the color scheme specified in the legend) and the genes detected for each taxon in the outer circle for MinION sequencing of the Shiga toxin-producing E. coli (STEC)-spiked beef sample after 1 h of sequencing.

The *stx* genes (*stx1* and *stx2*) were also linked to genomic regions of *Enterobacteriaceae* and to bacteriophages. The reads assigned to *Enterobacteriaceae* could also correspond to STEC bacteria, as *Enterobacteriaceae* is the family of the *Escherichia* genus. Shorter reads may not cover any species- or genus-specific genomic features, preventing their univocal assignment to a single higher level taxon. Such reads are attributed by BLAST to a common ancestor of higher taxa from which the read could potentially be derived, e.g., the family *Enterobacteriaceae*. The same could apply to reads classified as phages, as the *stx* genes present in the STEC genome derive from the integration of these phages, but these could also be present in their mobile form in the environment.

5.3.1.2. Outbreak Resolution and Strain Characterization From Long-Read Sequences by Strain-Level Inference, Gene Detection, and Single-Nucleotide Polymorphism Phylogeny

Finally, as an equivalent to the characterization of an isolate obtained in routine, a strain-level analysis was performed on all sequenced metagenomics reads to obtain clusters of reads corresponding to the different *E. coli* strains present in the sample. The presence of the STEC strain was confirmed based on the detection of *stx* genes in the clustered reads. It corresponded to a strain mapped to the taxon 741093 from the Metamaps analysis (RefSeq NC_017906.1, NCBI: txid741093). Two other non-pathogenic strains were detected in the samples and mapped to the Metamaps proprietary taxa x494 (RefSeq NZ_CP019271.1, NCBI: txid562) and 745156 (RefSeq NZ_CP009166.1, NCBI: txid745156) (Dilthey et al., 2019). Metamaps uses an extended database taxonomy where some NCBI taxonomic nodes are further subdivided to ensure higher resolution of taxonomic assignment. The same strains were detected after Illumina sequencing (Saltykova et al., 2020). The STEC strain was further investigated for SNP phylogeny to relate it to other cases (i.e., isolates from food and human origin related to the same outbreak as the spiked isolate and sporadic cases). Strains inferred from Illumina sequencing of the same sample (Buytaers et al., 2020) were also included in the tree (Figure 5.3). The inferred STEC strain obtained after 12, 24, and 48 h of MinION sequencing clustered with the corresponding isolates and metagenomics strain obtained from Illumina sequencing, with 0 SNPs distance (Supplementary Material 3), and separated from the sporadic cases. The presence of three virulence genes of importance for STEC characterization (*eae, stx1*, and *stx2*), as well as the serotyping genes (O-type and H-type), was also confirmed in the genome of the inferred STEC strain. The serotype and virulence genes in the inferred STEC strain correspond to the genes present in the strain that was spiked. The reference coverage from the MinION run starting from 12 h of sequencing was comparable to the coverage obtained from isolates of the same outbreak and strain inferred from Illumina metagenomic sequencing and therefore considered as sufficient for a phylogenetic analysis. Shorter sequencing time on the MinION did not offer sufficient coverage to conduct the phylogenetic analysis (Supplementary Material 3).

5.3.2. Investigation of Long-Read Flongle Sequencing as a Less Expensive Alternative for Strain-Level Metagenomics Outbreak Investigation

The same sample of beef meat containing an endogenous population of non-pathogenic *E. coli* and spiked with a STEC pathogen, previously characterized to the strain level after Illumina sequencing (Buytaers et al., 2020), was sequenced on a Flongle flow cell to investigate a less-expensive alternative. However, as the output of the Flongle is approximatively 10 times lower than the MinION, we also sequenced on the Flongle DNA for which the extraction involved host removal, previously sequenced on Illumina (Buytaers et al., 2020), in an attempt to increase the amount of reads linked to the microbial pathogen. The data analysis on the sequenced long reads was the same as the data analysis presented for the long reads

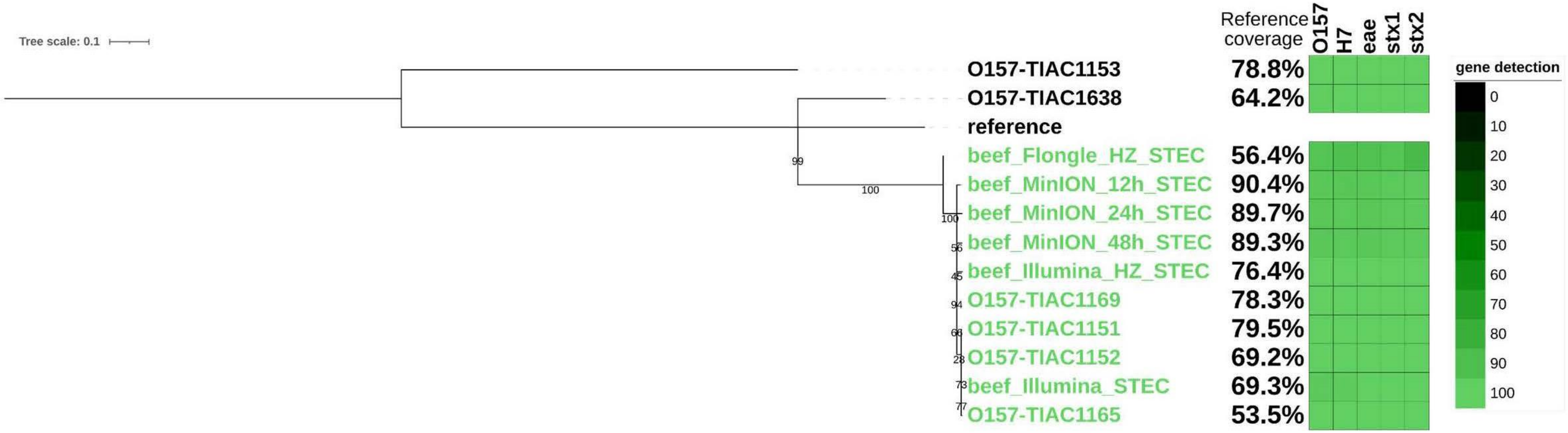


Figure 5.3: Single-nucleotide polymorphism (SNP)-based phylogenetic tree of STEC strains inferred from metagenomics sequencing (beef) and of sequenced isolates *with percentage of the reference genome covered (i.e., percentage of reference genome that is useful for SNP analysis, see section "Materials and Methods") and gene detection (O-type and H-type and genes eae, stx1, and stx2; green shaded blocks representing the query coverage) in each strain represented on the side of the branch. Isolates TIAC 1151, 1152, 1153, and 1638 are from food origin. Isolates TIAC 1165 and 1169 are from human origin. Reference: E. coli O157:H7 str. Sakai (BA000007.3). Green: closely related strains from the outbreak cluster. Black: sporadic cases outside the outbreak cluster. The scale bar represents nucleotide substitution per 100 nucleotide sites. Node values represent bootstrap support values.*

sequenced on the MinION. The analysis was also conducted at different time points of the Flongle sequencing run to determine the time needed to achieve the expected results.

5.3.2.1. Taxonomic Classification of All Sequenced Reads

After Flongle sequencing, the main genus detected in the sample without host DNA removal was *Bos* (Figure 5.1). The same bacterial taxa, with the exception of *Enterobacter*, were detected as for the Illumina and MinION sequencing, including *Escherichia*, but with a higher percentage of unclassified reads. As for MinION sequencing, a small portion (5%) of mammal reads were incorrectly classified as Ovis.

The DNA extract treated with host DNA removal agent (FlongleHZ) presented 2% of reads classified as *Bos* and 0.5% of reads classified as *Ovis*, although no reads were classified as mammals in the Illumina sequencing of the same DNA extract. However, this is a large decrease compared to the amount of Flongle reads classified as eukaryotes without the host DNA removal step (50%). The bacterial taxa detected were the same for this sample after Illumina or Flongle sequencing and, except for the absence of *Aeromonas* and *Comamonas* and the presence of Citrobacter and Lactobacillus, were identical to the bacterial taxa detected without host DNA removal. This difference might be explained by the presence of bacterial DNA in the extraction buffer or its presence at very low level in the food sample. *Escherichia* represented 10% of the reads, which is slightly higher than the values obtained without host DNA removal. The sample with host DNA removal sequenced on the Flongle presented the highest percentage of unclassified reads (39%).

5.3.2.2. Confirmation of the Presence of a Pathogen in the Flongle-Sequenced Metagenomics Sample Using in silico DNA Walking on Long Reads

Similar as for MinION sequencing, an *in silico* DNA walking was conducted in order to attribute a genomic context (taxon) to detected virulence genes. This analysis was conducted on all reads generated at different time points during the Flongle sequencing of the two DNA extracts (with or without host DNA removal).

After 1 h of sequencing, the virulence genes characteristic of a STEC (i.e., *stx, eae*, and *ehxA*) could be retrieved in the sample treated with host DNA removal (Figure 5.4A) and detected on genome fragments that could be assigned to *Escherichia*. Similarly, as with the MinION analysis, the virulence genes were also found associated in smaller proportions to *Enterobacteriaceae*, which correspond to the family of the *Escherichia* genus, or stx1 phage, the bacteriophage carrying the *stx1* gene that can be inserted in the STEC genome. The classification to a higher level (*Enterobacteriaceae* or phage) might be explained by the short length of the reads.

Without host DNA removal (Figure 5.4B), 5 h of sequencing were sufficient to obtain the required information to determine that the pathogen was present in the sample, i.e., virulence

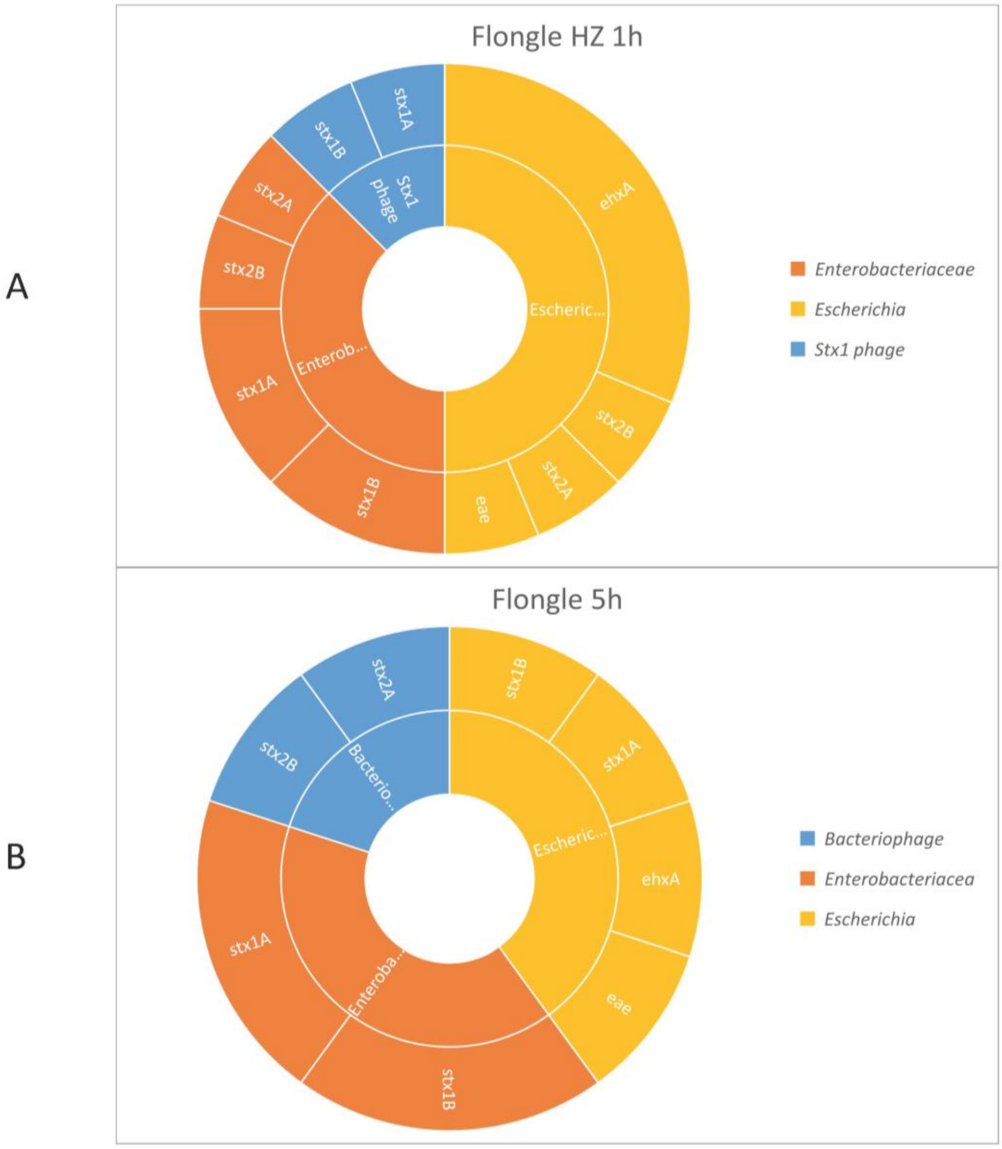


Figure 5.4: In silico DNA walking results, presenting the genera in the inner circle (following the color scheme specified in the legend) and the genes detected for each taxon in the outer circle of the STEC-spiked beef sample after DNA extraction with or without host removal. *(A) Flongle sequencing after 1 h of sequencing, DNA extract with host removal. (B) Flongle sequencing after 5 h of sequencing, DNA extract without host removal.*

genes *stx, eae*, and *ehxA* associated to *Escherichia* genome. Again, the virulence genes could be also assigned to *Enterobacteriaceae*, as well as bacteriophage for some *stx2* genes.

As a STEC is defined as an *E. coli* harboring an *stx* gene, the information presented was sufficient to conclude that a STEC was present in the samples, after 1 h of sequencing with host DNA removal and 5 h of sequencing without host DNA removal. However, a longer sequencing time would be required to obtain all virulence genes characterizing the strain that was spiked. In our workflow, the full characterization of the STEC present in the sample is done at the next step, after strain inference, in order to characterize specifically each potential pathogenic strain present in the sample.

5.3.2.3. Outbreak Resolution and Strain Characterization From Flongle Long-Read Sequences by Strain-Level Inference, Gene Detection, and Single-Nucleotide Polymorphism Phylogeny

The different *E. coli* strains present in the sample were inferred from the reads of the two Flongle sequencing runs, and the STEC strain was identified among these strains after detection of stx genes in the clustered reads (Supplementary Materials 4, 5).

The pathogenic strain corresponded to reads that mapped to the Metamaps taxon 741093 (RefSeq NC_017906.1, NCBI: txid741093) for the DNA extract without host DNA removal, i.e., a similar strain as found with MinION sequencing, and to Metamaps taxon x13 (RefSeq NZ_CP012802.1, NCBI:txid83334) for the DNA extract with host DNA removal, which is also a STEC O157:H7. The endogenous strains were mainly mapped as Metamaps taxon 745156 (RefSeq NZ_CP009166.1, NCBI: txid745156), similarly as for the MinION sequencing, as well as Metamaps taxon x311 (RefSeq NZ_CP019267.1, NCBI:txid562) for the extract without host DNA removal (Supplementary Materials 4, 5).

After SNP calling, it was observed that the coverage of the reference genome (Supplementary Material 3) was insufficient to conduct a SNP-level phylogenetic analysis (less than 1%) for the DNA extract without host DNA removal. Therefore, the inferred STEC strain obtained after Flongle sequencing of the DNA extract without host DNA removal was not included in the phylogenetic tree. However, 24 h of Flongle sequencing of the DNA extract with host DNA removal led to obtaining clustered reads covering 56% of the genome at or above 10× coverage, which was sufficient to cluster the metagenomics-derived strain with the outbreak cases on the phylogenetic tree (Figure 5.3). Serotyping genes (O-type and H-type) as well as virulence genes *eae, stx1*, and *stx2* could be detected with high identity in the strain, confirming that it was similar to the spiked strain. A distance of 0–3 SNPs per million genomic positions (Supplementary Material 3) was observed for the other isolates from the same outbreak as well as the metagenomics-derived strains from Illumina or MinION sequencing, which is in the expected range. However, the distances of the outbreak strain to the background isolates (TIAC 1153 and TIAC 1638) were somewhat lower with Flongle sequencing data after host DNA removal than with MinION and Illumina sequencing data (30 SNPs per million of genomic positions for the Flongle sequencing compared to 39–46 SNPs to TIAC 1638 for Illumina and MinION, and 126 SNPs per million of genomic positions for the

Flongle sequencing compared to 139–155 SNPs to TIAC 1153 for Illumina and MinION; Supplementary Material 3), indicating that not all SNPs could be called at the obtained coverage.

5.4. Discussion

The rapid and precise characterization of a pathogen during foodborne outbreak investigation, as well as the tracing back to the food source, is crucial to stop further spreading of the infections. Therefore, a metagenomics approach has been proposed as an alternative to the currently performed microbiological analyses requiring a not-always-straightforward isolation of the pathogenic strain. As previously described (Buytaers et al., 2020), Illumina sequencing may be used to obtain the full information necessary for outbreak investigation from metagenomics samples, without the need for isolation, and this to the strain level, after about one full week of lab work (Buytaers et al., 2021c). However, not only the need for more proofs of concept but also the high cost of such an analysis impact its potential implementation as a routine practice. To render the analysis somehow more cost-effective, while still taking the required coverage into account, 8–12 samples were pooled into one Illumina MiSeq run in previous studies (Leonard et al., 2016; Buytaers et al., 2020, 2021c). However, it might not always be possible to analyze this number of samples as the number of available food samples during outbreak investigation varies and is not gathered at a single time point. Besides, delaying the sequencing run to gather sufficient samples is not an option when a fast response is required, especially in outbreak investigation. Using a smaller number of samples in the run would however substantially increase the sequencing cost per sample. Moreover, these runs, generating 2×250-bp reads, have a set sequencing duration of 48 h, which is significant during ongoing outbreak investigations. Long-read sequencing and flexibility in sequencing time, which is made possible by ONT, could offer a solution to these drawbacks.

In this study, we first sequenced an artificially contaminated sample (beef containing an endogenous community of non-pathogenic bacteria including *E. coli*, spiked at very low dose with STEC), previously sequenced on Illumina (Buytaers et al., 2020), with a MinION flow cell. The data analysis followed the same flow as the analysis previously described for Illumina sequencing (Buytaers et al., 2020), but with adapted algorithms and tools for taxonomic classification, virulence gene detection, and genome inference of long reads. This allowed to match the contaminated food with human isolates from the same outbreak, after a shorter sequencing time. After only 12 h of sequencing, endogenous and pathogenic *E. coli* strains could be obtained from the sequenced reads, and the clustered reads corresponding to the STEC could be linked to outbreak isolates from food and human origin. The virulent strain-related reads harbored all virulence genes expected from the spiked bacteria and could be placed accurately in a phylogenetic tree with 0 SNP difference to the outbreak cluster, which is much lower than the SNP distance previously obtained after metagenomics ONT sequencing (Hyeon et al., 2018). The high number of SNPs observed by Hyeon et al. (2018) could be due to the specificity of the SNP calling procedure that was used. The authors applied a workflow based on the CFSAN pipeline, with a relatively low minimal allele frequency for SNP calling (0.6). These settings have, however, shown to exhibit a lower SNP calling accuracy even with the more accurate Illumina sequencing data, with higher SNP distances between the outbreak isolates as a result (Saltykova et al., 2018). Moreover, the long reads sequenced with ONT also offer the possibility to investigate at the same time the genomic context of reads carrying these virulence genes, in an *in silico* DNA walking approach. Recently, the European authorities proposed the attribution of the virulence genes to their respective bacterial host as an isolate-free alternative to confirm the presence of foodborne pathogens after metagenomics sequencing (EFSA, 2019b). We were able to detect the expected virulence genes (*stx, eae*, and *ehxA*) on genome fragments that could be linked to an *Escherichia* genome, therefore confirming the presence of the pathogen in the sample after only 1 h of sequencing on the MinION flow cell. This approach could eventually be implemented in real time while receiving data from the sequencer, as it has been shown previously for AMR genes (Leggett et al., 2020).

MinION sequencing offers the opportunity to work with long reads, allowing access to the genomic context of the reads sequenced, as well as the flexibility of real-time sequencing and sequencing one sample at a time. However, it remains an expensive consumable, and therefore, the lower cost Flongle flow cell was also tested. The Flongle flow cell was ideal to rapidly obtain a confirmation of the presence of a pathogen in the sample at the lowest cost after taxonomic classification and *in silico* DNA walking. Indeed, it allowed to confirm the presence of the STEC strain (detection of stx gene in *Escherichia* genome) after 1 h of sequencing if host DNA removal was conducted or 5 h with traditional DNA extraction. As the output of the Flongle flow cell is substantially lower compared to the output of the MinION flow cell, retrieval of information for strain comparison was only possible when host DNA removal was conducted during the DNA extraction. The coverage of the reference genome by the clustered reads corresponding to the STEC strain obtained from the extract without host DNA removal was not sufficient to establish phylogenetic links. The threshold to determine if a strain contains sufficient reads using metagenomics to perform further characterization or SNP phylogeny is hard to define strictly as lower coverages are also observed for genomics on isolated strains (e.g., TIAC 1165 covering 54% of the reference genome, Figure 5.3). More analyses such as the one within this study, including for other pathogens, are necessary to pinpoint such limits. We could also observe that the reference genome to which the reads were assigned after Flongle sequencing was different from the reference genomes mostly covered after MinION or Illumina analyses. However, the different references to which the STEC reads clustered were all STEC O157, and the interpretation of the results was not impacted by the reference (SNP-level phylogeny obtained for the different strains). A future alternative could be to pool reads assigned to groups of similar references instead of working with individual references, as already proposed in the work of Saltykova et al. (2020) for short-read sequences.

In the present work, a three-step analysis has been applied on food samples sequenced with different flow cells. For each step, the minimal time required to obtain results was assessed. First, a taxonomic classification to obtain an overview of the genetic content of the food sample, followed by virulence gene detection coupled to an *in silico* DNA walking method for hazard identification. These tests could be performed in a very fast timeframe of a few hours, depending on the treatment of the DNA extract and the selected sequencing flow cell, and could even be implemented in real time in the future. This could potentially solve partially, i.e., when food leftovers are available and it concerns a bacterial origin (for which isolation is currently the routine approach), the issue of foodborne outbreaks for which the food source cannot be determined, accounting for 60% of all (i.e., also including other agents as source) outbreaks notified in the EU (EFSA, 2021c). Finally, as a third step, strain-level phylogeny in order to relate human cases of an outbreak to its food source can be achieved after 12 h of sequencing on a MinION flow cell or 24 h on a Flongle flow cell if host DNA removal was applied during the DNA extraction. Notably, for *in silico* DNA walking, a threshold of one read harboring an stx gene and traced back to the *Escherichia* genus was considered as sufficient to determine the minimal time to suspect the presence of a pathogen in the sample. However, a discussion within the international scientific community is necessary to determine such threshold, and we recommend to continue the sequencing after this minimal time to collect more information. Moreover, obtaining and characterizing the pathogenic strain (third step) are still necessary to confirm the suspicion. Based on this work, a new strategy for detection of bacterial pathogens in food, using shotgun metagenomics, could be proposed to the reference laboratories (Figure 5.5): a screening of all food samples that might be related to a foodborne outbreak, including those for which the contaminant is unknown, for pathogens using the Flongle and taxonomic classification followed by virulence gene detection and *in silico* DNA walking, potentially in real time. This might involve additional enrichment media and/or conditions to be able to cover all bacterial foodborne pathogens, depending on the specific outbreak based on patient's symptoms, to fully replace the conventional way of working. Once the presence of a bacterial pathogen is confirmed in a food sample, this analysis can be followed by a strain-level read classification and phylogeny that can be attempted on the Flongle sequencing data or, if not possible, based on further Illumina or MinION sequencing. Also, the choice of the sequencing technology will depend on a cost analysis based on the amount of samples to sequence as well as the timeframe to obtain results and the capacities of the laboratory. This strategy could be run in parallel with attempting to obtain a bacterial isolate from the same food samples, which would ideally be sequenced with WGS in order to populate the still necessary databases that are required to perform the metagenomics-based bioinformatics analyses. The perspective of such a strategy consolidates the new perception that metagenomics has the ability to be used as a new alternative for outbreak investigation, source attribution, and risk assessment of foodborne microorganisms (EFSA, 2019b). In order to implement the same data analysis applied to artificially STEC-contaminated samples in this study to other bacterial pathogens, the same workflow can be followed, and only the databases for gene detection and read mapping have to be adapted according to the contaminant(s) detected through taxonomic classification.

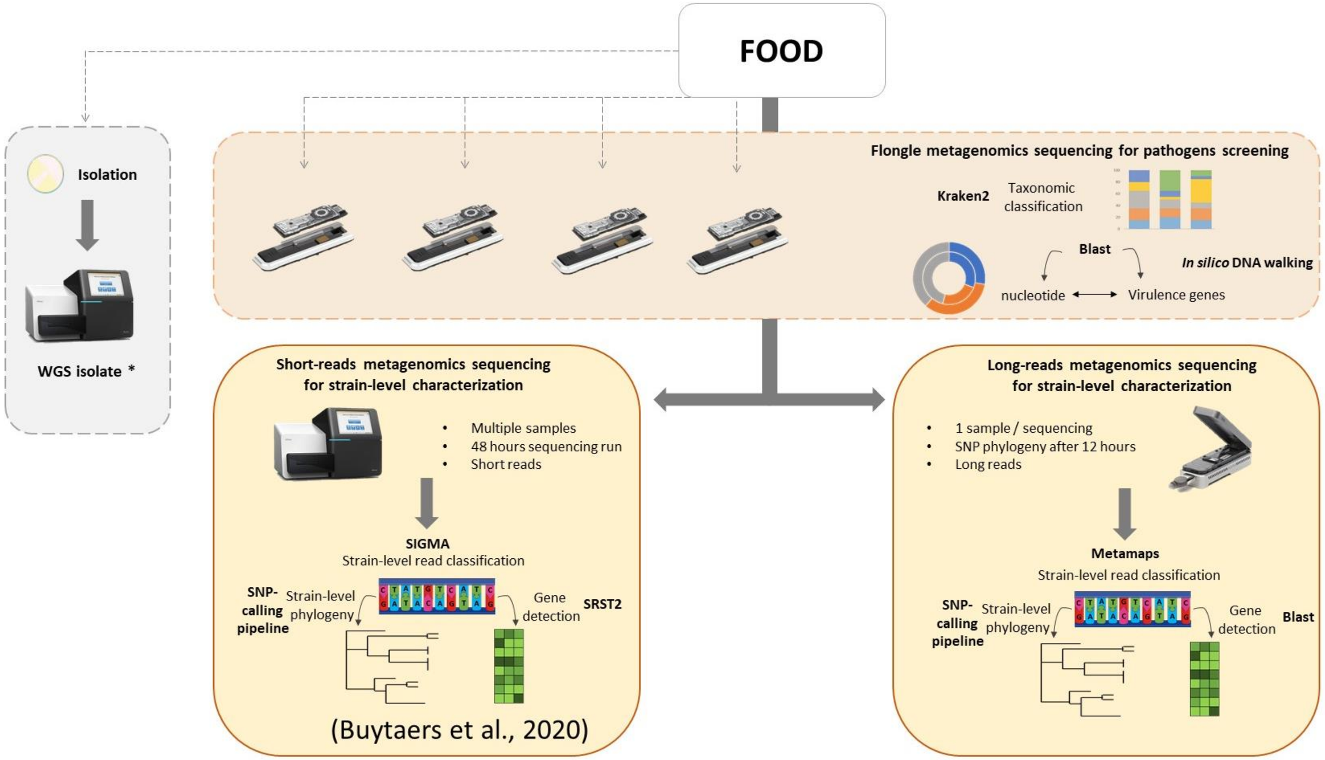


Figure 5.5: Integrated metagenomics-based strategy for microbiological foodborne outbreak investigation. *As first optional steps, food samples can be screened for the presence of bacterial pathogens using metagenomics Flongle sequencing, taxonomic classification, and in silico DNA walking (based on BLAST to the nucleotide database and virulence genes database) in parallel with the ongoing attempt to isolate the pathogen, followed by WGS of the obtained isolates. A strain-level characterization can be attempted from the Flongle sequencing or conducted after using an Illumina strategy (more cost-effective for multiple samples) or a MinION strategy (fast response for one sample) in food samples for which the presence of a pathogen is confirmed. The strain-level data analysis for Illumina sequencing was previously presented Buytaers et al. (2020). The strain-level data analysis workflow for MinION sequencing is based on classification using Metamaps, a gene detection with BLAST, and phylogenetics with a SNP-calling pipeline. The asterisk (*) indicates that WGS isolate data is interesting to feed to reference genome databases for the classification of the metagenomics reads for future analyses.*

The implementation of such a metagenomics approach in routine, however, still requires overcoming several challenges. First, the data analysis currently requires sufficient informatics hardware, especially performant GPUs for real-time base-calling and analysis. Additionally, trained bioinformaticians are needed, as no automated pipeline has been developed yet for a strain-level pathogen characterization. Benchmarking studies comparing more bioinformatics tools need to be performed to identify tools allowing to obtain similar results in the same, or even faster, timeframe. For this, we believe that studies such as this one offer interesting datasets to be explored further. Second, some consumables like the Flongle flow cells, which have a very short storage life, can be difficult to obtain in a short timeframe when the demand exceeds the production capacities as experienced during the COVID-19 pandemic. The output of Flongle flow cells might also be difficult to predict due to the possible instability of the very low amount of pores not only before loading but also after loading. This might affect low-level contamination samples. The MinION sequencing resulted in a larger output, allowing the potential applicability to other samples, regardless of the quality of the flow cell (number of pores). The work of Forghani et al. (Forghani et al., 2020) showed that a quasimetagenomics method to the strain level with MinION sequencing can be extended to other strains of STEC and other bacterial pathogens without a problem. However, our study, although only including one serotype, showed the potential of the metagenomics approach for samples presenting a population of several different *E. coli* strains (including non-pathogenic strains). More studies might be necessary to validate the potential of long-read strain-level metagenomics for food safety assessment and foodborne outbreak investigation for other pathogens, including viruses and parasites. The enrichment and extraction methods might have to be adapted depending on the pathogens to investigate. Moreover, while we analyzed samples contaminated with the lowest infectious dose, more studies with different contamination loads might lead to a more precise limit of detection for the method, especially as the number of pathogenic cells is undetermined after enrichment.

In conclusion, this work is a proof of concept of the potential to conduct real-time and affordable strain-level outbreak investigation based on ONT long reads, testing the potential of the MinION as well as the Flongle flow cells. Although a limited amount of samples and only one STEC strain was included in our proof of concept study, we demonstrated the ability to obtain the characterization and relatedness of a STEC spiked at a very low dose in a food matrix based on metagenomics sequencing on a MinION flow cell after only 12 h or on a Flongle flow cell after 24 h if host DNA removal was applied during the DNA extraction. Moreover, we also presented a rapid strategy to confirm the presence of a pathogen in a food matrix based on long-read sequencing without the need for isolation (i.e., *in silico* DNA walking). All this was possible on food samples enriched in a non-selective medium following the ISO practice. This makes it particularly interesting for reference laboratories when only limited quantities of the food samples are left, and there is no need for sequential culturing steps on pathogen-specific selective media, with pathogen-specific growth conditions. Moreover, with the method we propose, food that has been enriched at the reference laboratory can be sequenced with a metagenomics workflow in parallel to the isolation protocol, without the need for a different enrichment protocol, which can be particularly interesting when no isolate can be obtained. Finally, these results allowed proposing a more global perspective, as a metagenomics-based strategy to be used by the routine (reference) laboratories, determined by the required level of information required, cost-effectiveness, and timeframe to obtain results. This contributes to the demand of the EFSA asking to demonstrate the ability of metagenomics to be used as a new alternative for risk assessment, source attribution, and outbreak investigation (EFSA, 2019b).

Acknowledgements

We are grateful to Stefan Hoffman from Transversal Activities in Applied Genomics (Sciensano) for the initial set-up of Oxford Nanopore Technologies in the service.

Supplementary data

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.738284/full#supplementary-material

CHAPTER 6
A shotgun metagenomics approach to detect and characterize unauthorized genetically modified microorganisms in microbial fermentation products

This chapter was adapted from the previously published manuscript:
Buytaers, F.E.; Fraiture, M.-A.; Berbers, B.; Vandermassen, E.; Hoffman, S.; Papazova, N.; Vanneste, K.; Marchal, K.; Roosens, N.H.C.; De Keersmaecker, S.C.J. A Shotgun Metagenomics Approach to Detect and Characterize Unauthorized Genetically Modified Microorganisms in Microbial Fermentation Products. Food Chem. Mol. Sci. 2021, 2, 100023, doi:10.1016/j.fochms.2021.100023.

<u>Authors' contributions:</u>

F. E. Buytaers was responsible for data curation, formal analysis, investigation, methodology, validation, visualization and writing the original draft. M. Fraiture helped for the investigation, resources, review & editing of the manuscript. B. Berbers participated to the investigation, resources, visualization, review & editing. Els Vandermassen, S. Hoffman and N. Papazova took part in the resources, review & editing. K. Vanneste helped for the resources, software. K. Marchal participated in the supervision. N. H. C. Roosens took part in the conceptualization, formal analysis, funding acquisition, investigation, resources, validation. S. C. J. De Keersmaecker was involved in the conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, writing of the original draft. All authors have read and agreed to the published version of the manuscript.

Abstract:

The presence of a genetically modified microorganism (GMM) or its DNA, often harboring antimicrobial resistance (AMR) genes, in microbial fermentation products on the market is prohibited by European regulations. GMMs are currently screened for through qPCR assays targeting AMR genes and vectors, and then confirmed by targeting known specific GM constructs/events. However, when the GMM was not previously characterized and an isolate cannot be obtained, its presence cannot be proven. We present a metagenomics approach capable of delivering the proof of presence of a GMM in a microbial fermentation product, with characterization based on the detection of AMR genes and vectors, species and unnatural associations in the GMM genome. In our proof of concept study, this approach was performed on a case with a previously isolated and sequenced GMM, an unresolved case for which no isolate was obtained, and a non-GMM-contaminated sample, all representative for the possible scenarios to occur in routine setting. Both short and long read sequencing were used. This workflow paves the way for a strategy to detect and characterize unknown GMMs by enforcement laboratories.

6.1. Introduction

A GMO (genetically modified organism) is defined as "an organism in which the genetic material has been altered in a way that does not occur naturally by mating and/or natural recombination" (Article 2 of Directive 90/220/EEC). GMOs include transgenic animals, plants and genetically modified micro-organisms (GMM). Food and feed products including enzymes or additives (e.g. vitamins) are often produced through the use of GMMs to replace chemical synthesis methods, as this is more practical and requires less resources (Deckers et al., 2020a). A specific microbial species is chosen for its suitability and ease of cultivation, and is then genetically modified to produce, for instance, the required compound in large quantities through microbial fermentation (Deckers et al., 2020a). Selection of the genetically modified (GM) strain(s) is often conducted based on selective growth with antibiotics. To this end, antimicrobial resistance (AMR) genes are often inserted in the modified strain, e.g. in the expression vector used for introduction of the new characteristic. However, the introduction of full-length AMR genes in microorganisms that might end up in food/feed products, poses a potential public health risk. Indeed, these can be easily transferred to other species, including pathogens, thereby leading to treatment failure (WHO, 2018). In Europe, the GMM's authorization process for food enzymes and food additives (such as vitamins) falls under regulation EC 1331/2008, EC 1332/2008 and EC 1333/2008 (European Parliament and the Council of the European Union, 2008a, 2008b, 2008c). To allow that such microbial fermentation products can be produced with GMM in a contained environment, a confidential dossier must be submitted to the European commission, and EFSA performs a safety assessment. Moreover, these GMMs should be used only to produce microbial fermentation products and, in contrast to GM crop plants, are not intended for human and animal consumption. Therefore, these GMMs do not fall under EU regulations 1829/2003 and 1830/2003 for GMOs, and no dossier has to be submitted to the EU for authorization of the commercialization of GM food and feed for a specific GMM (European Parliament and the Council of the European Union, 2003a, 2003b). Consequently, the detection of an unexpected contamination by such a GMM in food and feed is per se unauthorized. This means that zero tolerance, including for its associated recombinant DNA, must be applied, i.e. neither viable cells nor DNA from the producer strain can be detected in the final commercialized product (Silano et al., 2019). Additionally, the companies producing these GMMs do not have to provide to the EU a method to identify them. This is in contrast with what is foreseen by regulations 1829/2003 and 1830/2003 for GM crop plants intended for the food and feed chain. Therefore, no detection/identification method for the GMM is available for enforcement laboratories. As several fermentation products were already shown to be contaminated by a living GMM or its DNA (Barbau-piednoir et al., 2015; Paracchini et al., 2017; Fraiture et al., 2020a), this calls for a proper control by enforcement laboratories.

Real-time polymerase chain reaction (qPCR) assays are the mandatory method for enforcement laboratories to screen and identify GM organisms (GMOs) in EU legislation 1829/2003 and 1830/2003. The aim is to ensure freedom of choice for the consumer by detecting unlabeled GMOs as well as the safety of food and feed. These qPCR assays target specific GM events, i.e. the insertion of the GM element(s) in the host genome which leads to unnatural associations. However, for GMM that are not intended for food and feed consumption, until recently no official methods were available that allowed their detection and characterization, as elaborated above. A novel strategy to detect and identify GMM in food and feed products was only recently developed. First, a qPCR screening is performed based on the detection of AMR genes and expression/shuttle vector carrying AMR genes. Hereto, a variety of qPCR tests have been developed, targeting the *cat*, (Fraiture et al., 2020c), *aadD* (Fraiture et al., 2020b) and *tet* (Fraiture et al., 2020d) genes, conferring a chloramphenicol, kanamycin and tetracycline resistance respectively, and targeting the shuttle vector pUB110 carrying *aadD* (Fraiture et al., 2020a). These qPCR tests can be complemented with conventional PCR methods followed by Sanger sequencing. This will allow obtaining additional information on the presence of microbial DNA and their species/genus identification, using for example 16S rRNA or ITS-based methods (Deckers et al., 2020c, 2020b), as well as on the presence of full-length AMR genes (Fraiture et al., 2021c). Demonstrating the presence of the full-length AMR gene is valuable information for risk assessment on the potential spread of this gene to other microorganisms in the environment including the human/animal gut after ingestion. If the screening based on AMR genes and/or vector is positive, thereby raising a strong suspicion of the presence of a GMM, a second line of analysis should be performed. This has the goal to both target unnatural associations (construct- or event-specific methods) and identify the GMM, thereby proving its presence in the microbial fermentation product (Barbau-piednoir et al., 2015; Paracchini et al., 2017; Fraiture et al., 2020a, 2020e). However, unlike it is the case for authorized GMOs as requested by regulations 1829/2003 and 1830/2003, no event/construct-specific method has been provided *a priori* by the producing companies to identify these GMMs.

If no second line qPCR analysis is available, the proof of the presence of a GMM can be obtained through whole genome sequencing (WGS) of a microbial isolate obtained from the fermentation product. This allows the identification of the unnatural association (Fraiture et al., 2020a). The knowledge of the DNA sequence can then lead to the future development and validation of a targeted GM-specific qPCR assay to be used in the identification step. This was the case for the identification method targeting a GM Bacillus overproducing vitamin B2 (RASFF2014. 1249) and one overproducing a protease (RASFF2019.3332) (Barbau-piednoir et al., 2015; Fraiture et al., 2020a). The GMM isolation process can however be arduous as the species and therefore the culture conditions are unknown. Moreover, isolation is not always possible, if the GMM is non-viable or non-culturable. Genetic modifications requiring the presence of a growth factor in order to culture the GMM are often encountered and unknown to the enforcement laboratories. Multiple species can also be present, and one of them can be missed by culturing. In other cases, only DNA of the GMM is present in the fermentation product. When no isolate can be obtained, a culture independent strategy has to be performed. For instance, a DNA walking method, as a targeted sequencing approach, can be used to detect unnatural associations. However, in order to apply this strategy, a minimum of knowledge is required. Indeed, the DNA walking strategy needs to anchor on a known sequence, like ARM genes and vector detected via the first line qPCR screening, in order to be able to characterize unknown flanking regions. Moreover, the DNA walking strategy can be time-consuming when regions of several kbps need to be covered, requiring successive DNA walking assays of each approximatively 1 or 2 kbps. A DNA walking strategy anchored on the pUB110 vector was previously used to identify the GM Bacillus overproducing alpha-amylase (RASFF2019.3332) (Fraiture et al., 2020e). Similarly as for the WGS approach, this can then subsequently lead to the design of new event/construct-specific methods (Fraiture et al., 2020e). Until now, WGS on isolates and DNA walking have enabled the development of qPCR methods allowing to identify 3 GMM constructs. However, in all the other scenarios, no fast and universal method is available to detect the presence of a GMM in a sample. This constitutes a major bottleneck for current GMM control, as many applications involving GMM are submitted to the European commission (for example, over a hundred dossiers for food enzymes mention the use of GMMs (Deckers et al., 2020a)).

A method not requiring prior isolation nor prior knowledge on the sequences, detecting all genes, including potential unnatural associations and potential species identification at once in a sample, would pave the way towards an open approach of generalized detection and characterization of unknown GMMs in microbial fermentation products. A shotgun metagenomics approach, i.e. sequencing all DNA from a sample, allows detection of any gene of interest as well as the detection of the species. It can also potentially reconstruct (partially) the genome of the strain(s) present, allowing to identify unnatural associations. This technology has been previously described for the successful characterization of food-borne pathogens at the strain level after a culture-based enrichment (Leonard et al., 2016; Buytaers et al., 2020). However, although the application to the field of GMM characterization is linked to a lower complexity of the microbial communities, some bottlenecks need to be addressed. First, not performing any enrichment is preferred to avoid the issue of species-specific growth conditions and non-viable GMM. Therefore, the shotgun metagenomics sequencing should be done in sufficient depth to allow for data analysis. This puts constraints on the cost-efficiency of the approach. Second, as the output of the metagenomics sequencing is a mix of reads representing all DNA present in the sample, putting the puzzle together to the species' genomic level, including detecting the unnatural association, is not straightforward. Short-read Illumina sequencing (max 300 bp reads), already described for metagenomics approaches in food using reference genome databases (Leonard et al., 2015; Buytaers et al., 2020), might not be sufficient for GMM, where this sequence information is largely missing. Long read sequencing such as offered by Oxford Nanopore Technologies (ONT) might facilitate the reconstruction of the genome (Somerville et al., 2018), especially with unnatural constructions such as GMMs. Moreover, it can help to obtain the unnatural associations or the full-length AMR gene, which are usually longer than 300 bp, on a single read. Several flow cells are currently on the market for this technology, e.g. the conventional MinION flow cell, and the newly released lower-output but also lower-cost Flongle flow cell, requiring half of the starting DNA material. As the metagenomics approach is still very expensive, the use of cheaper sequencing consumables such as the Flongle might contribute to reducing the price of the analysis while keeping a sufficient level of information (Grädel et al., 2019). However, ONT has been described to have a higher error rate as compared to short read sequencing, which could affect the results (Kono and Arakawa, 2019). This metagenomics approach has not yet been applied within the GMM field.

In this study, we present the first attempt to develop a strategy based on shotgun metagenomics for the general detection and characterization of GMMs in microbial fermentation products. Hereby, we envisaged to determine if and which AMR genes and shuttle vectors are present, and simultaneously provide information to identify the species present in the sample and to unequivocally prove the presence of the GMM by characterizing unnatural associations in its genome. To deliver a proof of concept of our approach, we have selected three samples, representative of the possible scenarios to occur in a routine setting, i.e. a previously analyzed sample containing a GMM *Bacillus subtilis* overproducing vitamin B2 (riboflavin), isolated and fully characterized at that time (RASFF 2014.1249) (Barbau-piednoir et al., 2015; Paracchini et al., 2017; Berbers et al., 2020), a sample positive for some qPCR markers but for which no isolate could be obtained and a sample with no GMM contamination. The short and long read sequencing technologies were compared for their performances, including the newly released Flongle, as a smaller and cost-effective alternative. The most appropriate data analysis workflow was considered, depending on the sample type and applied sequencing technology. This allowed to investigate the following hypothesis:

A shotgun metagenomics approach using short or long read sequencing is capable of detecting and (partially) characterizing unauthorized genetically modified microorganisms present in microbial fermentation products.

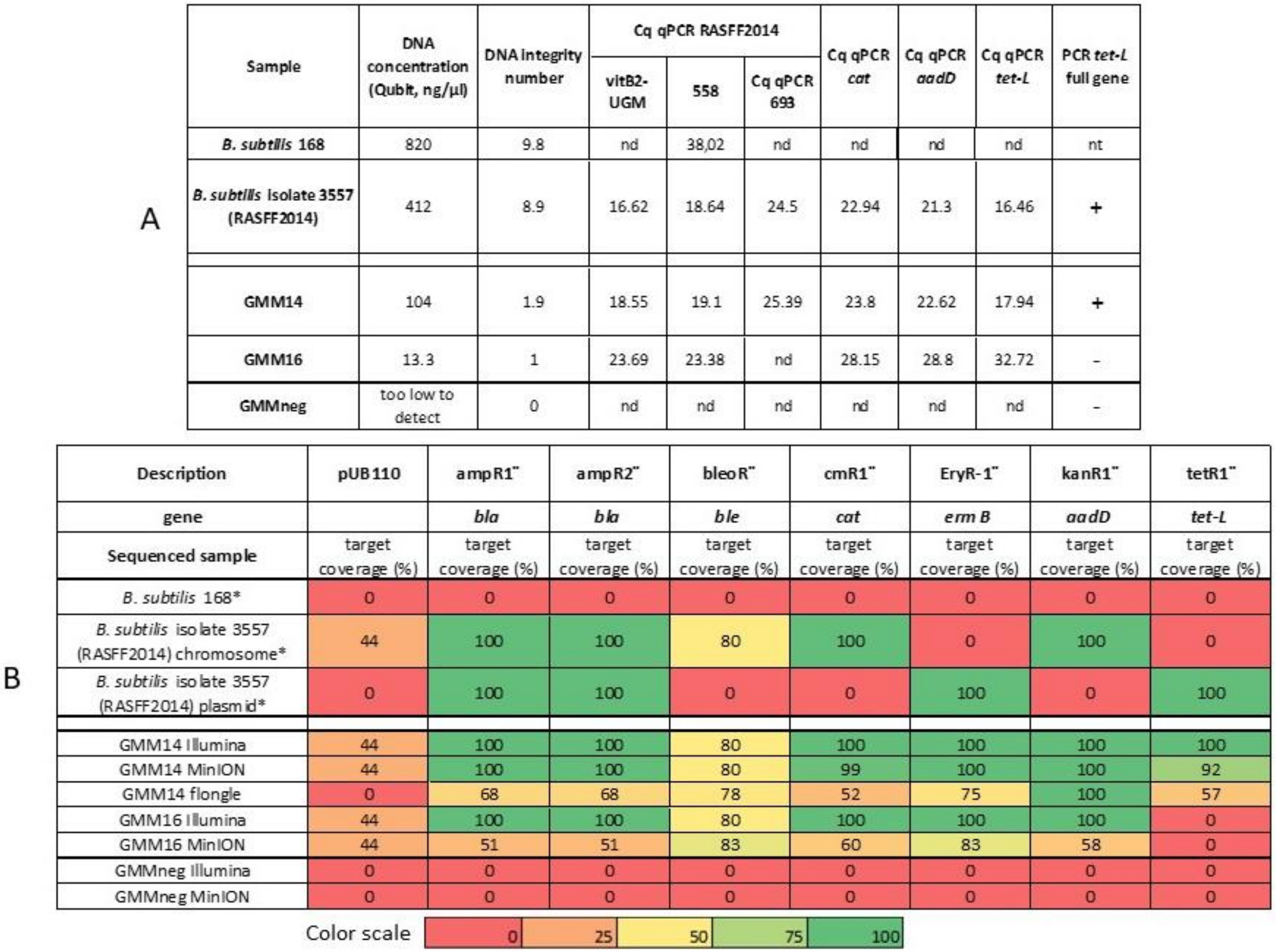
6.2. Materials and methods

6.2.1. DNA extraction and qPCR

Three samples of vitamin B2 (riboflavin) were investigated: one sample from 2014 containing a living GMM Bacillus strain (GMM14, RASFF 2014.1249), one sample containing DNA but negative to the previously developed qPCR methods (see paragraph below and Table 6.1) targeting AMR markers typical of GMMs and event-specific targets of the 2014 strain (GMMneg) and one sample from 2016 containing DNA corresponding to features of the 2014 strain (GMM16), but for which no strain could be isolated.

DNA was directly extracted from the vitamin powders without culture-based enrichment. Briefly, 200 mg of the sample was used for DNA extraction using the Nucleospin Food kit (Macherey-Nagel, Düren, Germany). The protocol was followed according to the manufacturer's instructions. qPCRs and PCR were performed on the DNA extracts as well as on DNA extracts from isolates of B. subtilis strains 3557 (GM) and 168 (wild-type), obtained during a previous study (Berbers et al., 2020) as described in Supplementary Materials 2.

Table 6.1: Characterization of GMM samples and bacterial isolates. A: DNA concentration and integrity, qPCR and PCR results B: detection results (AMR genes, pUB110) after isolate (168 and 3557) or metagenomics sequencing.



A

Sample	DNA concentration (Qubit, ng/µl)	DNA integrity number	Cq qPCR RASFF2014			Cq qPCR *cat*	Cq qPCR *aadD*	Cq qPCR *tet-L*	PCR *tet-L* full gene
			vitB2-UGM	558	Cq qPCR 693				
B. subtilis **168**	820	9.8	nd	38,02	nd	nd	nd	nd	nt
B. subtilis **isolate 3557 (RASFF2014)**	412	8.9	16.62	18.64	24.5	22.94	21.3	16.46	+
GMM14	104	1.9	18.55	19.1	25.39	23.8	22.62	17.94	+
GMM16	13.3	1	23.69	23.38	nd	28.15	28.8	32.72	-
GMMneg	too low to detect	0	nd	nd	nd	nd	nd	nd	-

B

Description	**pUB110**	**ampR1"**	**ampR2"**	**bleoR"**	**cmR1"**	**EryR-1"**	**kanR1"**	**tetR1"**
gene		***bla***	***bla***	***ble***	***cat***	***erm B***	***aadD***	***tet-L***
Sequenced sample	target coverage (%)	target coverage (%)	target coverage (%)	target coverage (%)	target coverage (%)	target coverage (%)	target coverage (%)	target coverage (%)
B. subtilis 168*	0	0	0	0	0	0	0	0
B. subtilis isolate 3557 (RASFF2014) chromosome*	44	100	100	80	100	0	100	0
B. subtilis isolate 3557 (RASFF2014) plasmid*	0	100	100	0	0	100	0	100
GMM14 Illumina	44	100	100	80	100	100	100	100
GMM14 MinION	44	100	100	80	99	100	100	92
GMM14 flongle	0	68	68	78	52	75	100	57
GMM16 Illumina	44	100	100	80	100	100	100	0
GMM16 MinION	44	51	51	83	60	83	58	0
GMMneg Illumina	0	0	0	0	0	0	0	0
GMMneg MinION	0	0	0	0	0	0	0	0

Color scale	0	25	50	75	100

A: DNA concentration (measured with Qubit); DNA Integrity Number of the DNA extracts (determined with Tapestation); qPCR detection of two junction sites specific to a GMM B. subtilis from RASFF 2014.1249 (VitB2_UGM and 558), a specific site in the plasmid (693) and three AMR genes (nd: not detected after 40 cycles); PCR of the full tet-L gene (located on the pGMrib plasmid, nt: not tested), B: Shuttle vector and AMR genes detection (¨, description based on list of common AMR genes detected in GMM from Fraiture, Deckers et al. (2020a)) in WGS data of the isolate of wild type B. subtilis strain 168 and GMM strain 3557 linked to RASFF2014 (, based on sequences from Berbers et al. (2020)) and in the assemblies from metagenomics sequencing using Illumina and MinION technologies, and reads from metagenomics Flongle sequencing of sample GMM14.*

Quality and quantity of all DNA extracts were evaluated using the Nanodrop 2000 (Thermo Fisher Scientific, Waltham, MA, USA), Qubit 3.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) and 4200 TapeStation (Agilent, Santa Clara, CA, USA). The latter results in the determination of a DIN (DNA Integrity Number) value representing the genomic DNA integrity based on fragment length (value between 1 and 10, with 10 reflecting the highest integrity).

6.2.2. Illumina (MiSEQ) shotgun metagenomics sequencing

The three DNA extracts were further processed using the Nextera XT library preparation kit (Illumina, San Diego, CA, USA) before sequencing on the Illumina MiSeq, generating paired-end 250-bp reads with the reagent kit v3 according to the manufacturer's instructions. The samples were sequenced in one run of 4 libraries (including another sample not belonging to this study), generating 2,895,502, 2,314,885 and 957 reads for GMM14, GMM16 and GMMneg, respectively.

6.2.3. Oxford nanopore technologies (MinION) shotgun metagenomics sequencing

The DNA libraries were prepared with the Genomic DNA by Ligation protocol (SQK-LSK109; Oxford Nanopore Technologies, Oxford, UK). The preparation was performed according to the recommendations for MinION sequencing, with one sample on one MinION flow cell. When the DNA concentration of the sample was too low to obtain 1 µg in 48 µl as recommended (samples GMM16 and GMMneg), 48 µl of the available DNA were used as input DNA. The prepared library was then loaded on a primed flow cell (R9.4.1) and a 48-hours sequencing run was started. The resulting fast5 files were basecalled using Guppy on fast mode (version 3.2.1, Oxford Nanopore Technologies), generating 971,569 reads with a median length of 302 bp for GMM14, 1,247,825 reads with a median length of 215 bp for GMM16 and 2,187 reads with a median length of 214 bp for GMMneg. The statistics were obtained using NanoPlot version 1.28.0 (De Coster et al., 2018) (full statistics in Supplementary Materials 1).

6.2.4. Illumina data analysis

The reads were trimmed using Trimmomatic version 0.38.0 with a sliding window approach requiring an average Phred score of 20 evaluated on a window of 4 bases (Bolger et al., 2014). Taxonomic classification of the reads was conducted using Kraken2 version 2.0.7 (Wood et al., 2019) as described in Buytaers et al. (Buytaers et al., 2020). The Illumina reads were assembled using SPAdes version 3.13.0 (Bankevich et al., 2012) with –meta mode. The presence of AMR genes in the contigs was detected using Blastn 2.7.1 on the ResFinder database (Kleinheinz et al., 2014) and on a set of the most common AMR genes detected in bacterial GMMs described in Fraiture, Deckers et al. (Fraiture et al., 2020b). For shuttle vectors detection, the database UniVec was first tested but this did not give satisfying results (no detection of the expected vectors, e.g. pUB110, described to be present in the isolate by Berbers et al. (Berbers et al., 2020), maybe due to incorrect metadata). Therefore, a Blast was only performed on the reference sequence of the pUB110 shuttle vector (GenBank: M19465.1) as it is well documented and described to be used in several GMMs including RASFF2014.1249 (Berbers et al., 2020; Fraiture et al., 2020e). This pUB110 vector is linked to the presence of the *aadD* AMR gene (Fraiture et al., 2020e). The presence of the riboflavin producing genes (rib operon of *B. subtilis* and *Bacillus amyloliquefaciens* as annotated with Prokka from the reference sequence from Berbers et al. (Berbers et al., 2020)) was also investigated with Blast. All Blast analyses were conducted with default parameters. Contigs that had a hit for AMR, pUB110 or rib genes were extracted and annotated using Prokka version 1.11 (Seemann, 2014) and then blasted online to the nucleotide database (https://blast.ncbi.nlm.nih.gov/Blast.cgi) to determine the species detected on the contigs. This was visualized using SeqBuilder Pro 15 v15.3.0 (DNASTAR Lasergene). Finally, the reads were mapped to the reference genome of the isolate from RASFF 2014.1249 (accession chromosome: NZ_CP045672.1 and plasmid: NZ_CP045673.1), using BWA MEM version 0.7.17 with default settings (Li and Durbin, 2010). The breadth of coverage to the full genome, chromosome and plasmid was calculated using SAMtools version 1.9 (Li et al., 2009) and awk (using the command: samtools depth -a alignment.sorted.bam | awk '{c++; if($3>0) total+=1}END{print (total/c)*100}') and the mapping was evaluated using QualiMap version 2.2.1 (Okonechnikov et al., 2015). The mapping was also visualized on IGV (Robinson et al., 2011) and the plasmid was manually annotated for the sites of the *tet-L* gene, the qPCR VitB2_UGM and the qPCR 693 following annotations from Berbers et al. (Berbers et al., 2020).

6.2.5. MinION data analysis

The fastq was converted to fasta format using Seqtk version 1.3 (https://github.com/lh3/seqtk/blob/master/README.md). The fasta file was used for taxonomic classification and species identification using Kraken2 and the same database and parameters as for Illumina reads. A more detailed species identification was conducted with Megablast using Blastn 2.7.1 (Camacho et al., 2009) to the regions V3-V4 of 16S rRNA sequences from Deckers, Vanneste, Winand, and Keersmaecker et al. (Deckers et al., 2020c) combined to the 16S rRNA database available on NCBI, as well as to the NCBI nucleotide database ((Sayers et al., 2019)with max_target-seqs set to 1. The fastq was used for assembly using Canu version 1.8 (Koren et al., 2017) modifying the parameters stopOnLowCoverage to 1, minReadLength to 200 and minOverlapLength to 100 in order to fit the relatively short reads obtained in the sequencing runs. Gene detection was conducted directly on the contigs using Blastn with the same parameters and on the same databases as for Illumina assembled reads. The contigs that had a hit were annotated using Prokka and blasted online to the nucleotide database to determine the species from which the sequences originated from. The mapping on the reference genome linked to RASFF 2014.1249 and calculation of the breadth of coverage were performed in the same way as for the Illumina data analysis (using -x ont2d command to use BWA MEM on ONT reads).

6.2.6. Flongle sequencing and data analysis

The DNA extract from sample GMM14 was also sequenced on a Flongle flow cell, after library preparation with the Genomic DNA by Ligation protocol (SQK-LSK109; Oxford Nanopore Technologies, Oxford, UK) following the manufacturer's recommendations. The library was loaded on a Flongle presenting 60 available pores at the start of a run of 24 h. The resulting fast5 files were basecalled using Guppy on fast mode (version 3.2.1, Oxford Nanopore Technologies) generating 60,093 reads with a median length of 306 bp.

The Flongle data analysis was similar to the MinION data analysis except for the assembly that could not be achieved due to low coverage. Therefore, the gene detection was conducted directly on the reads. As no contig was obtained, the search for unnatural associations was done as follows. The reads that had a hit for the presence of AMR genes were compared to the result of the Blast to the nucleotide database of the same read in order to obtain the species from which the sequence originated.

6.2.7. Data availability

All sequencing data is publicly available at NCBI SRA under project PRJNA686880.

6.3. Results

6.3.1. Development of a shotgun metagenomics-based approach for the characterization of a GMM

6.3.1.1. Sample selection and preparation

For the development of the shotgun metagenomics-based method for the characterization of GMM in microbial fermentation products, we selected a sample that was known to contain a GMM, and for which this GMM had been previously characterized after isolation. We selected the sample linked to the RASFF 2014.1249 (GMM14), containing a *B. subtilis* overproducing vitamin B2 (riboflavin) and therefore positive for the GM-events VitB2_UGM (Barbau-piednoir et al., 2015) and 558 (Paracchini et al., 2017). This GMM had been isolated and fully sequenced before (Berbers et al., 2020). As a negative control sample (GMMneg), we included a vitamin B2 sample from which DNA could be extracted but without detection of the GM-events specific to vitamin B2 overproduction (i.e. VitB2_UGM and 558). Therefore this sample was considered as 'probably not containing vitamin B2 overproducing GMM'.

To verify the samples, we performed a qPCR detection of event-specific markers as well as the presence of 3 AMR genes (*cat, aadD tet*) on the DNA extracts (Table 6.1.A). The markers specific to the GMM strain from RASFF 2014.1249 were detected in sample GMM14 (as expected), confirming the result of the initial screening. The 3 AMR genes were detected in GMM14, as expected based on the full genome sequence of the GMM *B. subtilis* that was previously isolated from this sample (Berbers et al., 2020). The same qPCR markers were detected in the DNA extracted from the isolate *B. subtilis* strain 3557 previously described in the context of the RASFF 2014.1249. These markers were not detected in DNA from the wild-type *B. subtilis* strain 168. Similarly, none of the genetic markers were detected in GMMneg.

The extracted DNA was then used for library preparation for shotgun metagenomics sequencing. We investigated both short read as well as long read sequencing technologies, especially as we assumed that long read sequencing would facilitate downstream data analysis to identify unnatural associations as well as full-length AMR genes. However, long read sequencing requires highly concentrated and high molecular weight DNA (Nanopore Protocol, 2019). The DNA concentration in sample GMMneg was too low according to the standards of MinION sequencing (i.e. need for 1 µg starting material in 48 µl). Moreover, the integrity of the extracted DNA was very low in all samples, as determined based on the obtained DIN value (less than 2, Table 6.1.A). The samples were nevertheless included in the downstream sequencing analysis.

6.3.1.2. Gene detection in assemblies from shotgun metagenomics

After sequencing, we looked for the presence of pUB110 and AMR genes in the contigs (Table 6.1.B, Supplementary Materials 2, Supplementary Materials 3), both from the short read as well as from the long read MinION sequencing output.

A part of the pUB110 shuttle vector (corresponding to the same portion as detected in the isolate from RASFF 2014.1249) as well as the genes linked to resistance to ampicillin (*bla*), bleomycin (*ble*, not present in the ResFinder database), chloramphenicol (*cat*), erythromycin (*ermB*), kanamycin & neomycin (*aadD*) and tetracyclin (*tet-L*) were detected in GMM14 after both short and long read metagenomics sequencing (Table 6.1.B). The shuttle vector and all these AMR genes have been previously described to be present on the chromosome and pGMrib plasmid of the isolate mentioned in RASFF 2014.1249 (Table 6.1.B, (Berbers et al., 2020)). Moreover, pUB110 is harboring the AMR genes *aadD* and *ble*. The assemblies of Illumina and MinION reads both allowed a detection of almost all genes with a coverage of more than 90%, except ble (with a coverage of 80%). The ble gene was already covered only at 80% in the chromosome of the isolate previously sequenced (Berbers et al., 2020), meaning that the recovery of what is expected to be present is 100%. The AMR genes were detected to cover the full-length of the reference genes on the contigs. However, the full-length genes were not detected on single reads, as the reads sequenced with MinION sequencing were shorter than the average length of an AMR gene. Nevertheless, the high coverage of the gene from the contigs is a strong indication that the full-length AMR gene was present in the sample.

As the GMM14 sample is a riboflavin feed additive, the presence of genes linked to riboflavin production (vitamin B2) was also investigated (results presented in Supplementary

Materials 2). The rib operon from *B. subtilis* and from *B. amyloliquefaciens* origin were both detected in the metagenomics sequencing of GMM14 with a coverage higher than 80% for the two sequencing instruments.

As expected, none of the most common AMR genes reported in GMMs were detected in GMMneg (Table 6.1, Supplementary Materials 2), confirming the negative results obtained with qPCR (Table 6.1). The riboflavin producing genes were also not detected in this sample after Illumina or MinION sequencing (Supplementary Materials 2).

6.3.1.3. Species identification via shotgun metagenomics

Next, we used the sequenced reads to identify the species present in the samples, to see whether these correspond to known microbial species used for known GMM based on patent information (Fraiture et al., 2020c). The sequenced reads from the two sequencing devices were classified per genus using Kraken (Fig. 6.1.A). *Bacillus* (green, Fig. 6.1.A.) was the main organism for the two sequencing methods in sample GMM14, although 35 to 40% of the reads could not be classified (light grey, Fig. 6.1.A.). More taxa were detected with Illumina sequencing but the small proportions might represent false positive classifications of some short reads. Two alternative methods aiming at obtaining more accurate information on the present species, were tested on the longer reads sequenced with the MinION: a Blast to a 16S rRNA database as shown in Fig. 6.1.B and a Blast to the NCBI nucleotide database as shown in Fig. 6.1.C. 786 of the MinION reads of GMM14 had 16S rRNA hits. With the two methods, *B. subtilis* could be detected as the main species in the sample (green, Fig. 6.1.B and C). However, other *Bacillus* species were sometimes detected with the 16S rRNA method (light green, Fig. 6.1.B). This could be expected as the 16S rRNA genes are very similar for these species and the method has been reported to be unable to differentiate efficiently between *B. subtilis* and *B. licheniformis* (Deckers et al., 2020c). The classification using the NCBI nucleotide database covers the full genome of each species, and thereby allows for more genomic markers to be used to attain species resolution. 31% of the reads could be classified to genus *Bacillus* sp. and 55% of all classified reads were detected as *B. subtilis* without ambiguity. The small proportion of *Escherichia* (yellow, Fig. 6.1.A, 6.1.B, 6.1.C) detected in both sequencing runs is partly explained by a misclassification (i.e. 12% of the reads classified as *E. coli* are mapping to the *B. subtilis* GM reference defined by Berbers et al. (Berbers et al., 2020)) but could also indicate the presence of DNA of this species in the sample. In conclusion, *B. subtilis* was detected in high proportions, corresponding to the GMM species that was previously isolated from the GMM14 sample. In the GMMneg sample, 62% of the reads were unclassified after Kraken analysis of the Illumina sequencing while *Bosea, Sphingomonas* and *Rhizobium* were detected as the main genera (Fig. 6.1.A). The latter two genera are known as common contaminants of Illumina sequencing (Winand et al., 2019). For MinION sequencing, more than 93% of the reads could not be classified, while *Ochrobactrum* and *Microbacterium* each represented 3% of the reads. *Bosea, Rhizobium* and *Ochrobactrum* are all part of the order *Rhizobiales*. The presence of these genera was not confirmed with the Blast to the 16S rRNA or nucleotide database (Fig. 6.1.B and 6.1.C). Indeed, no 16S rRNA hit was obtained in the MinION

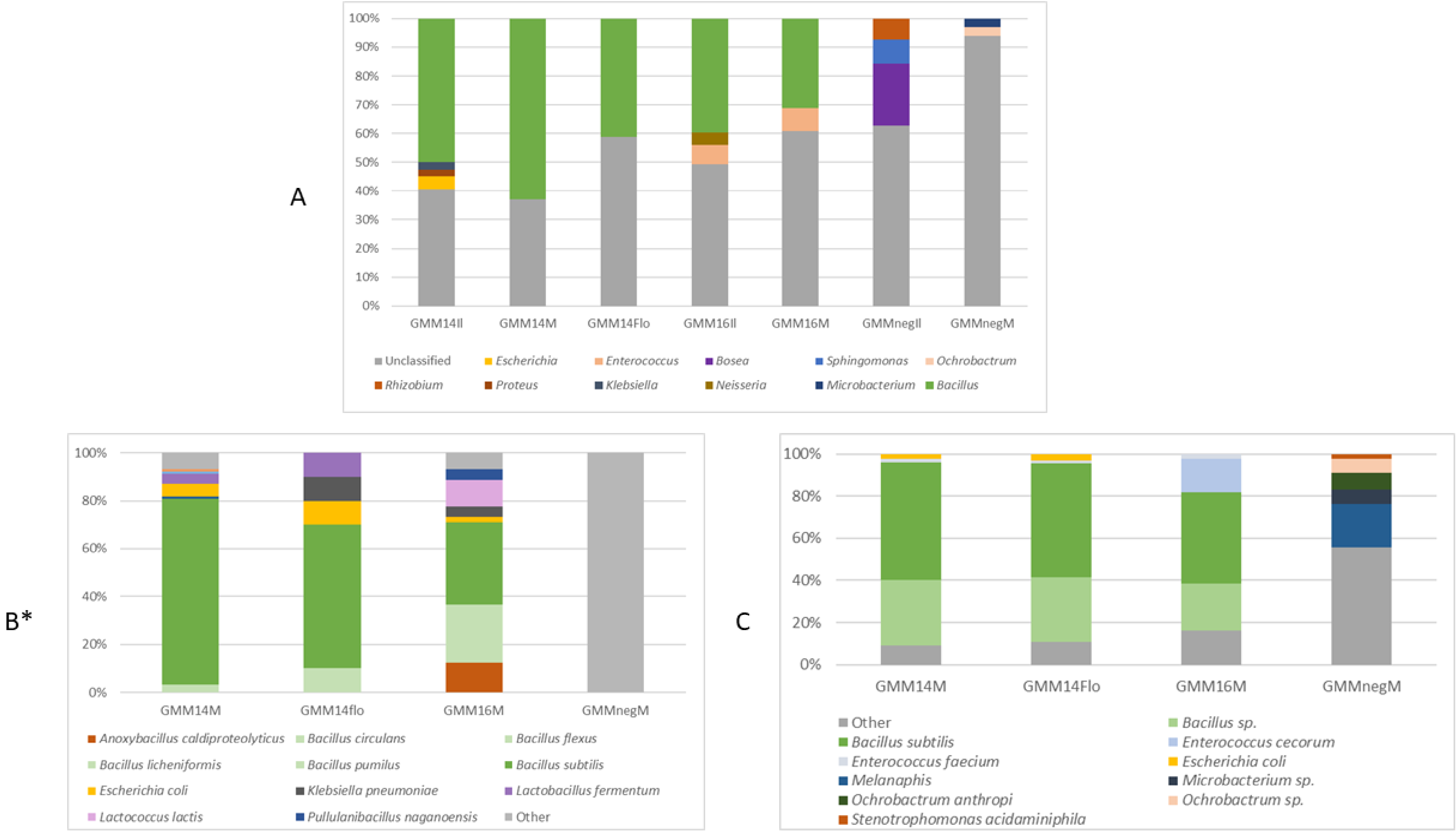

A
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
GMM14II
GMM14M
GMM14Flo
GMM16II
GMM16M
GMMnegII
GMMnegM
Unclassified
Escherichia
Enterococcus
Bosea
Sphingomonas
Ochrobactrum
Rhizobium
Proteus
Klebsiella
Neisseria
Microbacterium
Bacillus
B*
100%
80%
60%
40%
20%
0%
GMM14M
GMM14flo
GMM16M
GMMnegM
Anoxybacillus caldiproteolyticus
Bacillus circulans
Bacillus flexus
Bacillus licheniformis
Bacillus pumilus
Bacillus subtilis
Escherichia coli
Klebsiella pneumoniae
Lactobacillus fermentum
Lactococcus lactis
Pullulanibacillus naganoensis
Other
C
100%
80%
60%
40%
20%
0%
GMM14M
GMM14Flo
GMM16M
GMMnegM
Other
Bacillus sp.
Bacillus subtilis
Enterococcus cecorum
Enterococcus faecium
Escherichia coli
Melanaphis
Microbacterium sp.
Ochrobactrum anthropi
Ochrobactrum sp.
Stenotrophomonas acidaminiphila

Fig. 6.1: Species identification in the different samples. A: Kraken taxonomic classification results for Illumina ('Il'), MinION ('M') and Flongle ('Flo') reads. *Taxa representing <2% of the reads are counted in unclassified. B: Blast to 16S rRNA database results for MinION ('M') and Flongle ('Flo') reads. "Other" (grey): species representing <2% of the reads with hits (or for GMMneg: no hit obtained) *: Results presented to species level, as output from workflow described in Materials and methods section, however it was reported that 16S rRNA analysis is limited to genus level (Winand et al., 2019) C: Blast to nucleotide database results for MinION ('M') and Flongle ('Flo') reads. "Other" (grey): Species representing less than 2% of the reads with hits (e.g. Streptococcus pyogenes).*

sequencing of the sample, with the database used, while *Melanaphis, Microbacterium sp., Ochrobactrum anthropic* or *sp.* and *Stenotrophomonas acidaminiphila* were detected with the NCBI nucleotide database. The very low concentration of DNA in the sample (Table 6.1) led to a very low quantity of reads after sequencing, which could explain the inconsistency in genus identification for both sequencing methods. None of these genera are known as previously reported GMM (Fraiture et al., 2020c), and most probably represent a contamination. It should be noted that most likely in routine analysis, based on the negative results of the first line screening, no additional analysis would be performed. In our study, the sample was only used as a negative control for the metagenomics approach.

6.3.1.4. Detection of unnatural associations in the assembled metagenomic reads

The contigs containing AMR genes obtained for sample GMM14 were further investigated to determine if some unnatural associations (i.e. presence of parts of sequences belonging to different species or vector(s) in the same genome) were present. Given the nature of the sample, the same was done for contigs containing genes linked to riboflavin production (in this case the rib operon). As *B. subtilis* was detected as the main species, contigs harboring *B. subtilis* genome and parts of genomes from other species were investigated as probable unnatural association linked to the GMM in the sample (Fig. 6.2). This was done for the Illumina and MinION assemblies. Notably, several similar hits for genome, plasmid or vector identity could be obtained with the same confidence for the contigs investigated. However, only one hit was shown per region in the figure, in order to illustrate the unnatural association without aiming at identifying the exact origin of this segment of genome. An insertion of the chloramphenicol resistance gene (*cat*) in the genome of *B. subtilis* was detected with both sequencing technologies, interrupting the sequence of the *recA* gene (Fig. 6.2.A and 6.2.D). This corresponds to the 558 qPCR assay specific GM-event previously described for the RASFF2014 strain (Paracchini et al., 2017; Berbers et al., 2020). Another contig in the Illumina-based assembly contained 2 genes of the rib operon in the *B. subtilis* genome adjacent to a plasmid sequence originating from *Streptococcus pyogenes* (Fig. 6.2.B). Moreover, a part of the *B. subtilis* genome carrying *ribA* from the rib operon was linked to a part of an expression vector and the pUB110 plasmid sequence from *Staphylococcus aureus*, harboring 2 AMR genes (*ble* and *aadD*) in a contig from MinION sequencing (Fig. 6.2.C). The same pattern was observed in the chromosome of the GMM isolate described by Berbers et al. (Berbers et al., 2020). These sequences prove an unnatural association in the genome of *B. subtilis*, detected as the main species in the sample, and hence the presence of a GMM in sample GMM14.

As no AMR or *rib* genes were detected in GMMneg, this analysis was not conducted for this sample. This sample is considered not to contain a GMM strain.

6.3.1.5. Validation of method: Mapping of metagenomics reads to a previously characterized GMM reference genome

As a validation step to demonstrate that our metagenomics analysis detected the GMM previously characterized as an isolate from the same sample, we mapped the sequenced reads to the reference genome of this rapid alert (GCA_009914705.1 (Berbers et al., 2020)). We could show that the reference genome is fully covered with our metagenomics reads. The breadth of coverage was calculated as 100% for the chromosome and pGMrib plasmid with the two sequencing technologies, with a mean coverage of 57 on the chromosome sequence and 317 on the plasmid sequence for the MinION sequencing, and a mean coverage of 119 on the chromosome sequence and 492 on the plasmid sequence for the Illumina sequencing (Supplementary Materials 4). This additional validation step proves that the GMM detected with the metagenomics approach is indeed similar in genome structure to this previously sequenced isolate. This was expected since the sample analyzed originated from the same riboflavin product. Moreover, the mapping to the pGMrib plasmid was visualized (Fig. 6.3.A and B) with tags annotating the positions of the *tet-L* resistance gene (Berbers et al., 2020) and the positions for the qPCR VitB2_UGM (Barbau-piednoir et al., 2015) as well as the qPCR 693 (Paracchini et al., 2017) assay. These were all covered as expected from the qPCR results (Table 6.1).

6.3.1.6. Evaluation of Flongle sequencing

Based on the results described above, a workflow using either short or long read sequencing seems to enable the characterization of a GMM in a microbial fermentation product. However, in the short read sequencing run, more than one sample was included, to make it cost-effective. This might not be desirable in a routine set-up, where samples are often arriving, and hence need to be analyzed, on a one-by-one basis. The long read sequencing included one sample per flow cell, thereby rendering the cost per analyzed sample more expensive than the short read sequencing. Therefore, a Flongle sequencing was carried out on the GMM14 sample, as a less expensive, more flexible (one sample, with less input material required than MinION) and fast (24 h) sequencing alternative. The same analysis steps were performed as for the Illumina and MinION sequencing. This was done to evaluate whether the same information could be obtained using a long read sequencing device with a lower output (6% of the amount of reads for the same sample compared to the MinION sequencing, Supplementary Materials 1). All expected genes could be detected in the Flongle reads, but with a coverage starting from 52%, which would be filtered out of most classical analyses, and generally lower sequence similarity (80–90%) with a coverage starting from 52%, which would be filtered out of most classical analyses, and generally lower sequence similarity (80–90%)

to the reference sequence compared to the results obtained with the MinION contigs (>90%, Table 6.1, Supplementary Materials 2). The lower coverage is explained by the use of reads instead of an assembly. The use of short sequences could also explain why the shuttle vector pUB110 could not be detected with a coverage higher than 0.2% while 44% was present

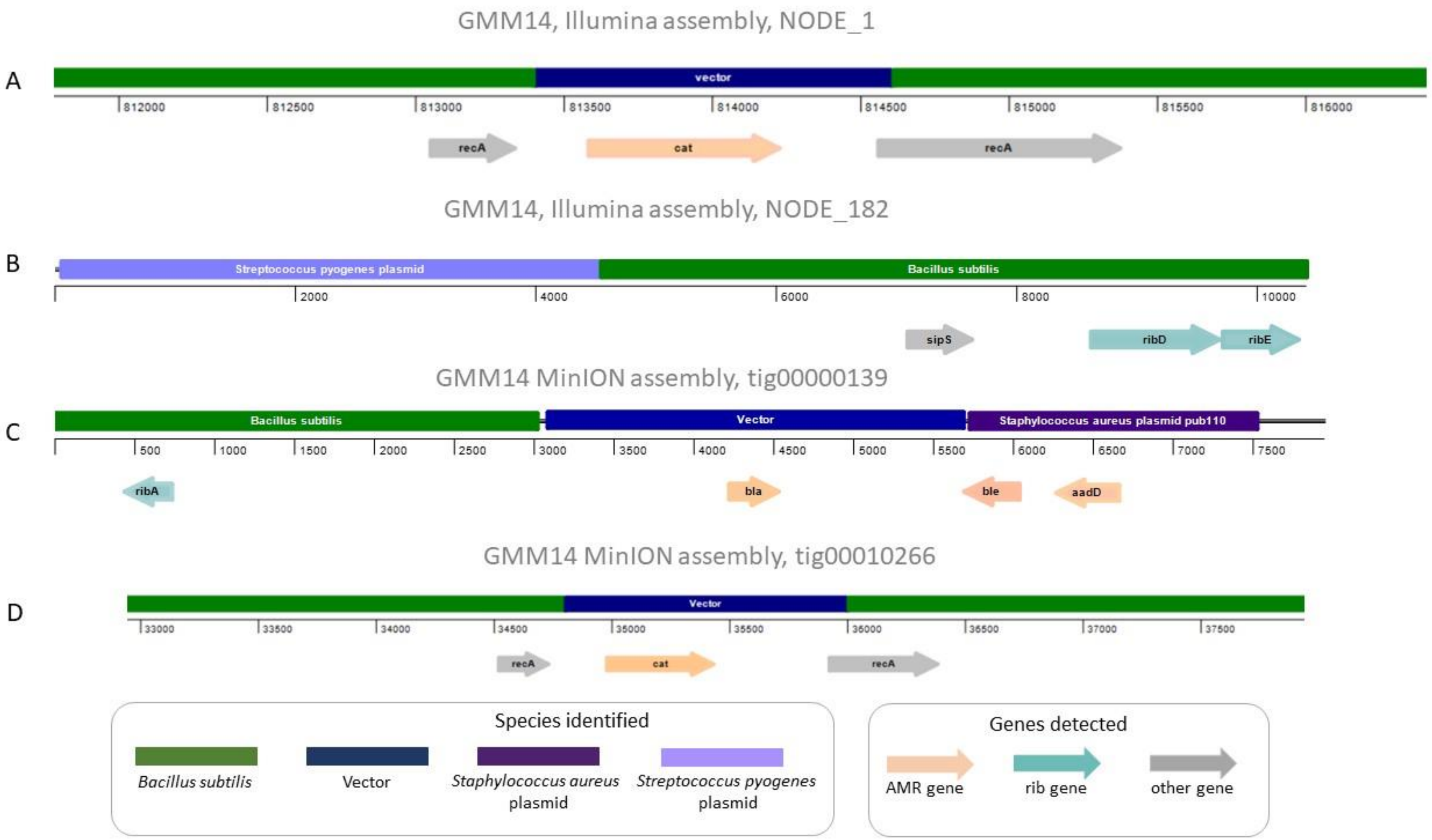


Fig. 6.2: Detection of species and genes on contigs of GMM14 sequenced with different technologies representing unnatural associations in the genome. *A-B: contigs from Illumina assembly. C-D: Contigs from MinION assembly.*

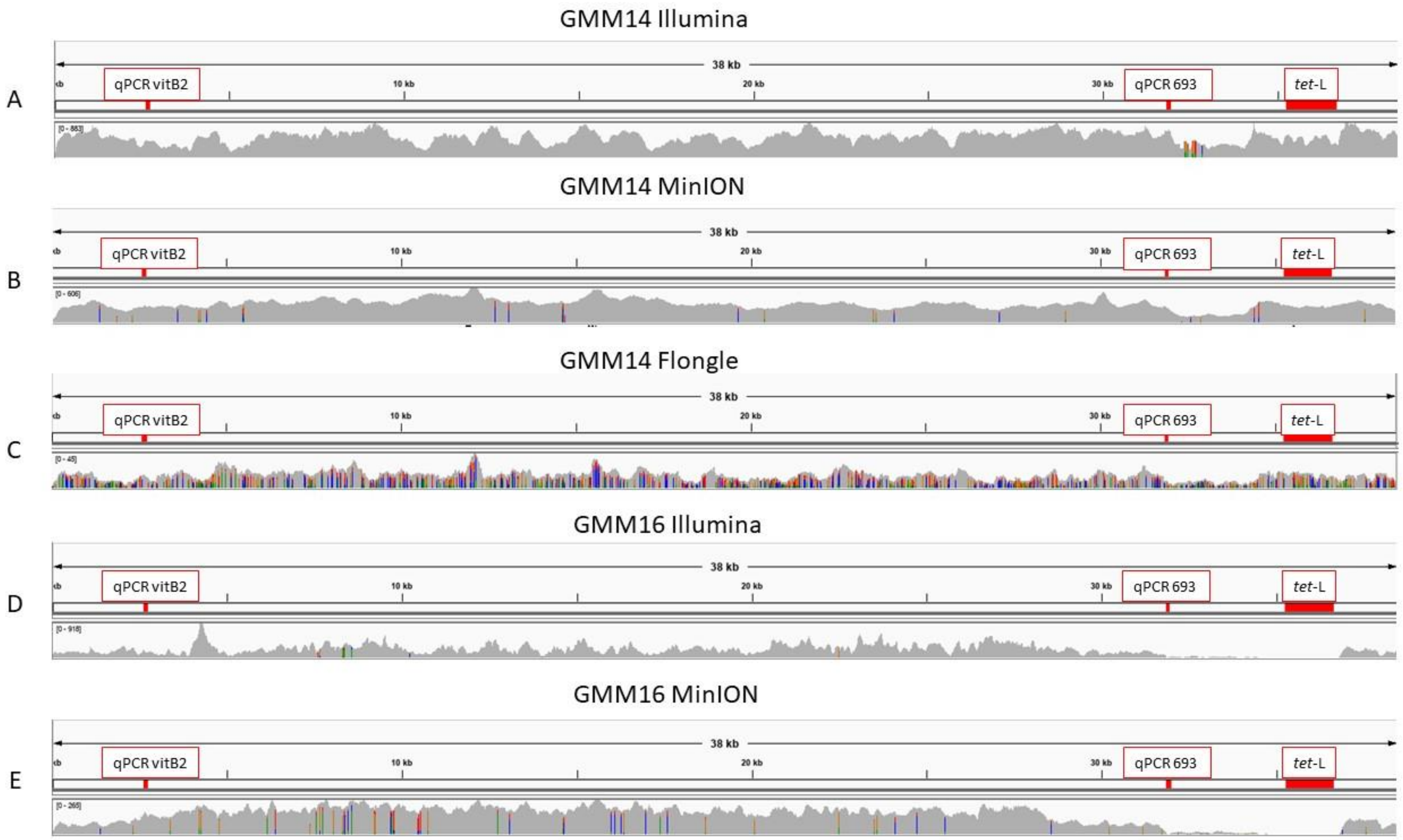


Fig. 6.3: Coverage of the pGMrib plasmid from* B. subtilis *strain 3557 (RASFF 2014) with annotation of the qPCR 693, qPCR VitB2 site and the* tet-*L gene. Colored bars: deviations from the reference. A: GMM14 Illumina sequencing. B: GMM14 MinION sequencing. C: GMM14 Flongle sequencing. D: GMM16 Illumina sequencing. E: GMM16 MinION sequencing.

in the reference genome from the isolate. Nevertheless, the AMR genes present on this plasmid (*ble* and *aadD*) were correctly covered. The genus *Bacillus* and the species *B. subtilis* were detected as the main microorganism in the sample, in the same proportions as for the Illumina and MinION sequencing (Fig. 6.1.A, B and C; 16S rRNA classification based on 31 reads with hits). As no assembly was obtained, the unnatural associations could not be detected as such. Nevertheless, an analysis with Blast to the nucleotide database of the reads that had a hit to an AMR gene confirmed that these AMR genes were detected in species or synthetic constructs other than *B. subtilis* (Supplementary Materials 5), the main species detected in the sample. This raises the suspicion about a possible alteration of the genome of *B. subtilis* to add AMR genes naturally present in other species. A mapping was performed to the reference genome of the isolate from RASFF 2014.1249 (Berbers et al., 2020) to confirm that it has a similar genome structure, as the same AMR and *rib* genes were detected. The reference genome was not fully mapped (92.6% breadth of coverage to the chromosome and 100% to the plasmid). A mean coverage of 3 was determined to the chromosome sequence and 15 to the plasmid sequence (Supplementary Materials 4). The low coverage, linked to the lower output of the Flongle, might explain the loss in breadth of coverage compared to the Illumina and MinION sequencing of the same sample. However, the obtained results indicated that the GMM detected using shotgun metagenomics Flongle sequencing of the GMM14 sample is similar in genome content to the previously sequenced isolate from RASFF 2014.1249. The mapping to the plasmid reference sequence was visualized as well (Fig. 6.3.C), indicating that the *tet-L* gene as well as the qPCR VitB2_UGM and qPCR 693 sites were covered, which corresponds the qPCR results (Table 6.1).

6.3.2. Applicability of the method: sample positive for GMM B. subtilis qPCR markers but without isolated bacterium

A vitamin B2 sample received for routine analysis in 2016 (GMM16), which tested positive for the GM-associated junctions VitB2_UGM and 558 by qPCR, but for which no living bacterium could be isolated, was used to demonstrate the applicability of our developed workflow. The re-extracted DNA gave, as expected, a positive qPCR signal for the vitamin B2 specific GM-events (VitB2_UGM and 558) and also for the 3 AMR genes (*cat, aadD* and *tet*) (Table 6.1). *tet-L*, which is known to be present on the pGMrib plasmid of the previously described GM *B. subtilis* (Berbers et al., 2020), was detected with a higher Cq of 32.7 compared to the 2 other AMR genes (*cat* and *aadD*), present on the chromosome of the same GM strain. This Cq was also higher compared to another qPCR marker that should be present on the pGMrib plasmid of the reference, the VitB2_UGM (Cq of 23.69, Table 6.1). The qPCR 693 assay (Paracchini et al., 2017), targeting the junction of pGMBsub03 to pUC19 located on the pGMrib plasmid in the GM *B. subtilis* isolate (Paracchini et al., 2017; Berbers et al., 2020), was not detected in this sample after 40 cycles of the assay. As this was a sign of difference with the previously described isolate from RASFF 2014.1249, a PCR of the *tet* gene (Fraiture et al., 2020d) was then performed on all samples to verify the presence of the full tet gene. This PCR was negative for GMM16 (Table 6.1). The DIN value of the obtained DNA extract was very low, indicating the presence of degraded DNA. This and the low concentration of the DNA (Table 6.1) were not optimal according to Oxford Nanopore's guidelines for MinION sequencing. Nevertheless, the DNA was used for short (Illumina) and long read (MinION) sequencing. MinION was selected over Flongle sequencing to account for the higher Cq value obtained for the detection of the tet marker, and hence anticipating a need for higher coverage/output.

6.3.2.1. Gene detection in assemblies from shotgun metagenomics sequencing

After gene detection in the assemblies from Illumina and MinION sequencing, the shuttle vector pUB110 and the resistance genes *bla, ble, cat, ermB* and *aadD* could be detected (Table 6.1.B). The shuttle vector was covered at 44% as observed for sample GMM14 and for the isolate from that sample. Most AMR genes were detected in full-length (100% target coverage) in the Illumina assembly except for ble, but the same percentage was covered as previously observed for sample GMM14 and the associated isolate. The assembly of the MinION reads allowed the detection of the same genes albeit with a lower coverage (the lowest being 51%) and a lower identity (see Supplementary Materials 2). Again, the genes were fully covered in the contigs, but the full-length genes could not be detected directly in the reads due to their limited length.

The genes detected were the same as the genes present on the previously characterized GMM14 isolate (and metagenomics GMM14 sample), except for the absence of the tetracycline resistance gene (*tet-L*) in the contigs, for which a higher Cq was obtained with qPCR.

Genes linked to riboflavin production were detected in assemblies from both types of sequencing reads for this sample (Supplementary Materials 2), i.e. genes from the rib operon from *B. subtilis* and *B. amyloliquefaciens*, confirming that most probably the DNA sequenced belonged to the organism producing the substrate. The coverage and identity of the detected genes was higher for the Illumina contigs than for the MinION sequencing, for which some genes were detected with a coverage lower than 50%.

6.3.2.2. Species identification via shotgun metagenomics

After taxonomic classification with Kraken (Fig. 6.1.A), more than 50% of the reads from the GMM16 sample could not be classified for both the Illumina and ONT data. The majority of the classified reads was attributed to the *Bacillus* genus after Illumina or MinION sequencing. This genus is listed as one of the most commonly referenced GMMs (Fraiture et al., 2020c), especially for the production of riboflavin. *Enterococcus* and *Neisseria* were detected in smaller proportions with the two sequencing technologies. Identification to a higher resolution was attempted with a Blast to the 16S rRNA database (Fig. 6.1.B, based on 596 reads with hits) and nucleotide database (Fig. 6.1.C). *B. subtilis* was detected at 34 and

43% with those methods, while other *Bacillus* species (for the 16S rRNA) or *Bacillus sp.* (for nucleotide) covered a remaining 20%. The other detected species were not consistent between the methods, and the presence of *Neisseria* was not confirmed.

6.3.2.3. Detection of unnatural associations in the assembled metagenomic reads

We looked for unnatural associations in contigs containing AMR genes and genes linked to riboflavin production (Fig. 6.4). A cat insertion in the sequence of the *recA* gene of *B. subtilis* was detected in the Illumina assembly of the sample (Fig. 6.4.A). The same insertion was described in the GMM linked to RASFF 2014.1249 (558 junction) (Berbers et al., 2020). Other unnatural associations of the genome of *B. subtilis*, harboring genes from the rib operon with plasmids from other species, were also detected in the Illumina and MinION assemblies (Fig. 6.4.B and 6.4.D). Moreover, an association of the *B. subtilis* genome, a vector containing two AMR genes (*aadD* and *ble*), and the *B. amyloliquefaciens* genome harboring the *rib* operon, was also detected in the MinION assembly (Fig. 6.4.C). The presence of these unnatural associations in the genome of *B. subtilis*, detected as the main species in the sample, proved the presence of a GMM in sample GMM16.

6.3.2.4. Mapping to a previously characterized GMM reference genome

Following the high similarity of the information detected in GMM16 with the previously characterized isolate from RASFF 2014.1249, except for the absence of the *tet-L* gene previously described to be present on the pGMrib plasmid of the GMM (Berbers et al., 2020), we conducted a mapping of the GMM16 metagenomics reads to the reference genome obtained for the isolate linked to RASFF 2014.1249 (Berbers et al., 2020). This resulted in a full mapping of the chromosome (100% breadth of coverage for the MinION sequencing and 99.9% for Illumina, Supplementary Materials 4) with a mean coverage of 24 after MinION sequencing and 39 after Illumina sequencing, but a partial mapping of the plasmid (99.9% breadth of coverage for the MinION sequencing and 97.5% for Illumina) with a mean coverage of 138 after MinION sequencing and 194 after Illumina sequencing (Supplementary Materials 4). A visualization of the mapping to the plasmid sequence showed the absence of reads mapping to the region of the *tet-L* gene (position 35241-36617 (Berbers et al., 2020)) in the metagenomics reads (Fig. 6.3.D and 6.3.E), in contrast to the metagenomics reads obtained for sample GMM14 (Fig. 6.3.A, 6.3.B and 6.3.C). This corroborates the absence of amplification of the full tet gene with PCR (Table 6.1). When zooming in (not represented in the figure), the region of qPCR 693 is also missing, confirming the result obtained with qPCR as well, while the rest of the pGMsub03, the region in which the qPCR 693 and the *tet-L* gene are described to be present in the reference strain from RASFF 2014.1249 (Berbers et al., 2020), is covered with very few reads, and not covered anymore if filtering for reads that map uniquely. All other regions of the pGMrib plasmid, however, were covered with reads. Therefore, we can conclude that the GMM present in the GMM16 sample, for which no isolate could be

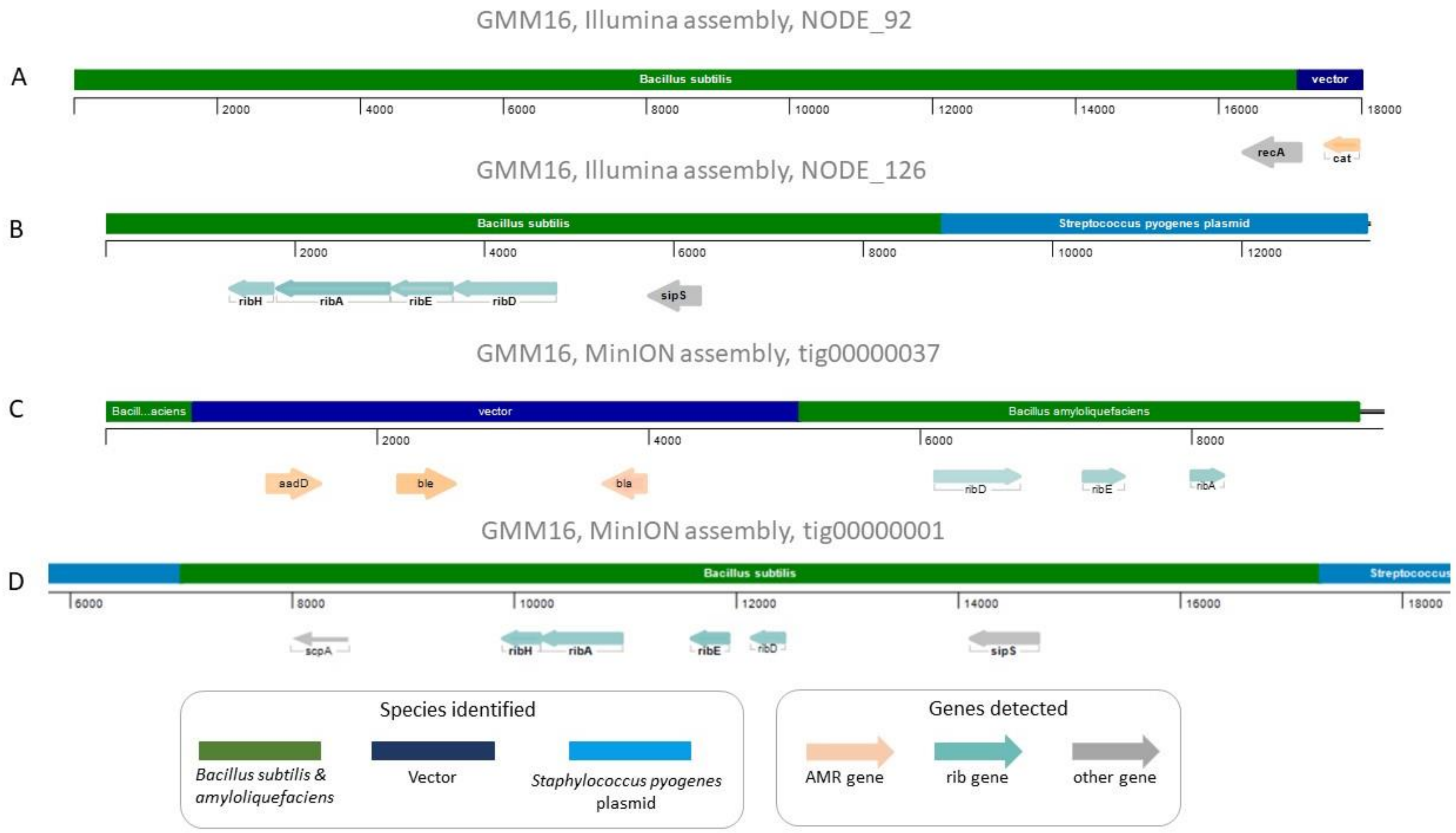


Fig. 6.4: Detection of species and genes on contigs and reads of GMM16 sequenced with different technologies to represent the presence of unnatural associations in the genome. *A-B: Contigs from Illumina assembly. C-D: Contigs from MinION assembly.*

obtained, is similar in genomic content at least for the chromosome to the isolate from RASFF 2014.1249. The plasmid might be different or have been modified but the region of the qPCR Vitb2_UGM as well as the erythromycin resistance gene were still detected with qPCR and/or after sequencing.

6.4. Discussion

GMMs are commonly used to produce microbial fermentation products. According to European regulation, the viable GMM or its DNA, often containing AMR genes, cannot be present in the final product of commercialized genetically modified food and feed (Deckers et al., 2020a). It is important for enforcement laboratories within Europe to have access to methods allowing the detection and characterization of such GMMs or their DNA. Construct/event-specific qPCRs have been previously developed on a case-by-case basis after WGS of an isolate (Barbau-piednoir et al., 2015; Fraiture, Bogaerts, et al., 2020; Paracchini et al., 2017. These methods have been used as a second-line analysis after detection of AMR genes and a shuttle vector for which first line qPCR assays have been developed based on publicly available patent information (Fraiture et al., 2020b). The development of such construct/event-specific methods, however, requires prior isolation of the contaminant for its characterization, and each test is only specific to one GMM. An alternative targeted approach, based on DNA walking, has been proposed and does not rely on obtaining an isolate (Fraiture et al., 2021c). However, it still requires prior knowledge to design primers and can be very laborious as the unnatural association might only be obtained after several consecutive reactions that each have to be carefully designed. The DNA walking approach has led to the design of an additional event-specific marker (Fraiture et al., 2020e). Using WGS or DNA walking methods, until now only 3 GMMs have been characterized and can be identified using qPCR. In this study, we propose an alternative open approach based on shotgun metagenomics to potentially allow untargeted identification of GMMs. This does not require isolation and allows detecting any AMR gene present in the DNA, identify the species present in the sample and expose the presence of unnatural associations of sequences in the genome. Our workflow was established with the aim to be usable in the future by the European enforcement laboratories as an alternative or addition to their current investigation tools.

Our results deliver a proof of concept for a shotgun metagenomics approach as a viable alternative to detect and characterize a GMM present in microbial fermentation products without the need for isolation or enrichment. In our workflow, the prediction of the presence of a GMM was based on the simultaneous detection of AMR genes or vectors in species previously described as common GMM producers (Fraiture et al., 2020c), and the encounter of unnatural associations in the genome (Fig. 6.5).

Altogether, our method allowed to achieve the same information as obtained with the currently used standard methods (detection of AMR genes or vectors with qPCR and detection of unnatural associations with WGS or with event-specific qPCRs). Moreover, it can potentially replace additional testing such as the detection of the genus/species with 16S rRNA-based methods. However, our method is able to perform all these analyses at once, thereby saving time. It even extends the characterization of the GMM, such as detecting the presence of AMR genes for which no qPCR methods have yet been developed (in this case *bla, ble, erm*), and identify to species level, even when multiple species are present, which is not always possible with the 16S rRNA method (Yang et al., 2016). This method also allows to describe previously unknown unnatural associations that could lead to the development of new event-specific qPCR methods.

We compared two sequencing technologies producing short reads or long reads. The results obtained with Illumina and MinION sequencing were equally satisfying, leading to the detection of all genes of interest and unnatural associations, with equal breadth of coverage after mapping to a reference genome. A shuttle vector and several unnatural AMR genes could be detected in the assemblies. This is a strong indication that the full-length gene is present in the samples. The identification of the reads to the species level was only obtained with the MinION sequencing with a Blast of the reads to the NCBI nucleotide database. This could be expected since the use of 16 s rRNA genes was previously described as insufficient to obtain species resolution (Winand et al., 2019). Moreover, the error rate of MinION sequencing is higher and might lead to a misclassification on the short and highly similar 16S rRNA region. This analysis could not be conducted on the short Illumina reads, however, illustrating the advantage of long read sequencing for species identification. The classification of contigs was not feasible for this application due to difficult or even dangerous interpretation of the results as by nature of the sample, these contigs represent an association of several species. Flongle sequencing yielded 6% of the amount of reads obtained from MinION sequencing, with a lower cost. These reads were of similar median length as the reads obtained with the classical MinION flow cell, but with a generally lower read quality, and allowed species identification and detection of the genes of interest. However, as no assembly could be performed, genes were detected with a lower coverage that might not pass classical thresholds of analysis. Although no thresholds for metagenomics analyses have been established yet, EFSA recently published a statement on the requirements for WGS analysis of isolated microorganisms intentionally used in the food chain (EFSA, 2021a). They advised query sequence hits with at least 70% length of the subject sequence to be reported when submitting a characterization dossier. Not being able to assemble the reads also complicated detection of unnatural associations. The breadth of coverage of the mapping was also lower due to some missing information in the lower output. Nevertheless, all information needed to prove that a GMM was present in the sample and to characterize it was obtained. Therefore, this is an interesting rapid and low-cost alternative for enforcement laboratories to get an overview of the content of a sample for which no information can be obtained with the normal qPCR screening, when the DNA concentration and quality are sufficient.

We have studied a previously described sample (GMM14) and then used the developed method to characterize a sample in which some GMM-specific markers were positive (qPCR VitB2_UGM, 558, detection of AMR genes) but for which no isolate could be obtained (GMM16). After thorough analysis of the reads and contigs obtained from this

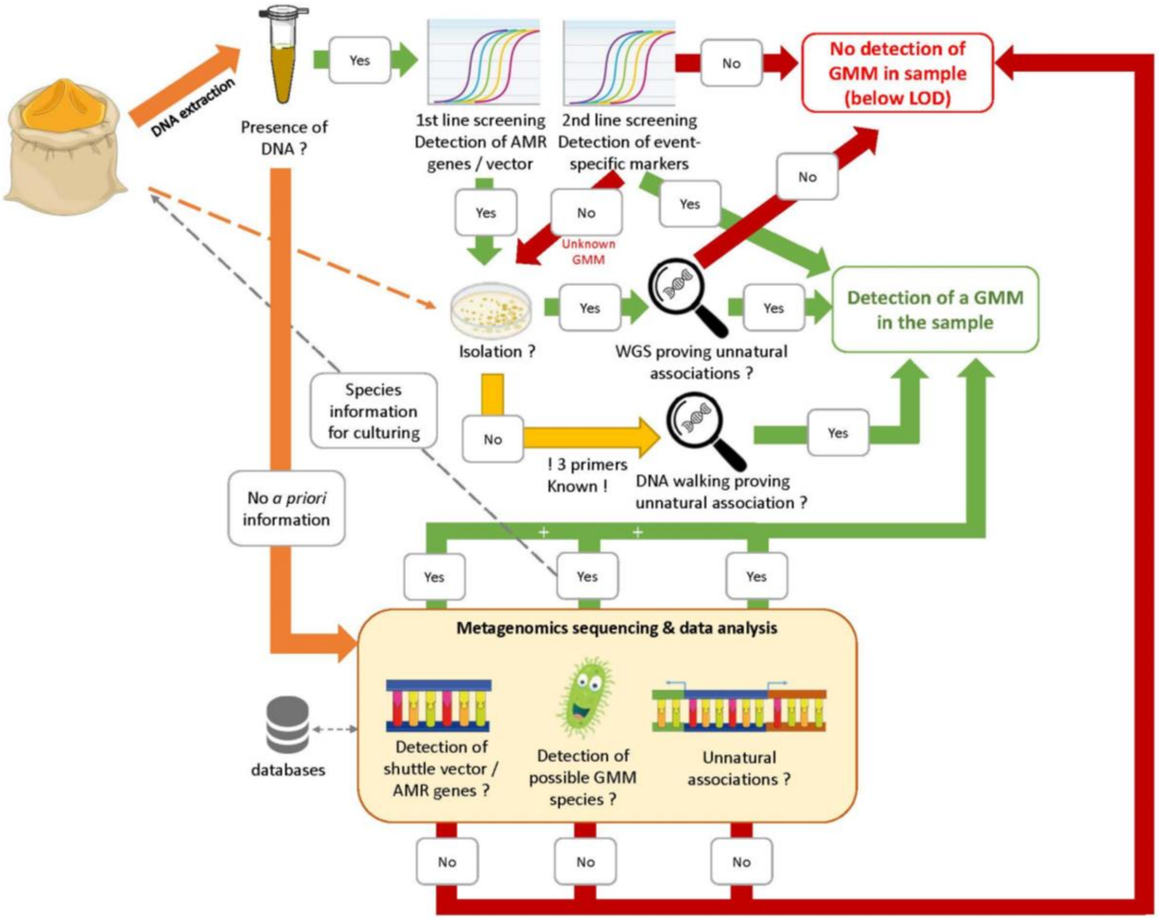

DNA extraction
Presence of DNA ?
Yes
1st line screening Detection of AMR genes / vector
2nd line screening Detection of event-specific markers
No
No detection of GMM in sample (below LOD)
Yes
No
Unknown GMM
Yes
No
Isolation ?
Yes
WGS proving unnatural associations ?
Yes
Detection of a GMM in the sample
Species information for culturing
No
! 3 primers Known !
DNA walking proving unnatural association ?
Yes
No a priori information
+
+
Yes
Yes
Yes
Metagenomics sequencing & data analysis
databases
Detection of shuttle vector / AMR genes ?
Detection of possible GMM species ?
Unnatural associations ?
No
No
No

Fig. 6.5: GMM detection decision tree, presenting the conventional workflow currently performed in enforcement laboratories (qPCR screening, DNA walking or WGS on the isolate) and the proposed metagenomics alternative when no isolate can be obtained. *If simultaneously detecting AMR gene(s) typically not naturally occurring, possible GMM species and unnatural associations in the genome, depending on the available databases, we can conclude that a GMM was detected in the sample.*

sample, we were able to detect, with both sequencing technologies, more AMR genes than detected with the qPCR (presence of *bla, ble, ermB*). We were also able to identify the main species as *B. subtilis* and detect unnatural associations in the genome, confirming that it was indeed a GMM. These parameters led to a strong suspicion that a similar GMM as the one previously characterized from RASFF 2014.1249 was present in this sample. The *tet* gene described to be present in the pGMrib plasmid of that GMM was however not detected after shotgun metagenomics sequencing of GMM16. A high Cq was obtained in qPCR screening, targeting a part of the *tet-L* gene and we could not demonstrate the presence of the full-length *tet-L* gene in the sample by PCR. After mapping to the reference genome (*B. subtilis* 3557, GenBank GCA_009914705.1, (Berbers et al., 2020)), we could establish that that part of the pGMrib plasmid was missing while the rest of the chromosome and the plasmid were fully covered by the sequenced reads. This suggests that this sample contains a similar but different GMM, with a plasmid that does not harbor the *tet-L* gene.

Our study is rather explorative. It needs to be seen as a proof of concept for the use of metagenomics approaches for the detection and identification of GMM. We illustrated this potential using a selection of samples representative for the possible scenarios in routine. In the future, additional samples need to be investigated. Moreover, some challenges still have to be overcome to make our workflow easier to implement in enforcement laboratories. First, short and long read sequencing both independently delivered the required result, demonstrating the presence of a GMM. Nevertheless, long read sequencing has some advantages in terms of costs, flexibility and species identification. However, the long read sequencing was performed with DNA extracts that were of lower quantity and quality, resulting in rather short median read lengths. If high molecular weight DNA could be obtained from the food/feed samples, the long read sequencing method could be used without the possible bias of assembly that can create chimeras. Moreover, the possible unnatural associations as well as full-length AMR genes might be detected on one (a few) single read(s). This would represent unequivocal proof of the presence of an AMR gene in the sample, potentially transmissible. For some MinION sequencing runs, the amount of DNA used in this study was not sufficient. Increasing the amount of sample material as input for the extraction could be a solution. Another alternative could be the enrichment of the sample by culture, maybe driven by information on which culture conditions to apply based on prior 16S rRNA analysis, in order to increase the DNA yield, and hence the flow cell output. These improvements could also pave the way towards a broader use of the Flongle flow cell if sufficient DNA quality and quantity can be obtained. It should be highlighted that during our study, the demand for the Flongle exceeded production capacities, especially with the needs for the current SARS-Cov-2 pandemic, leading to long waiting times aggravated by short expiration times that currently cannot match routine lab operating times. Besides, the treatment of the food/feed sample to remove viable cells and DNA as required by EU regulations, might have led to short fragments of damaged DNA already before its extraction. This would impede the possibility of extraction of high molecular weight DNA or the GMM to be enriched anyway. Therefore, although an assembly-free long-read based data analysis workflow would be ideal for unbiased detection of AMR genes, vectors and unnatural associations, the nature of the sample might force the use of assembly-based methods to identify a GMM. Second, the data analysis methodology we proposed is based on easy to use and well-established bioinformatics tools (Kraken, Spades, Blast, etc.). However, the development of push-button bioinformatics pipelines would be needed to allow full implementation in enforcement laboratories. Indeed, although next or third generation sequencers could be present in official control labs, the bioinformatics expertise for the application of these analyses might be missing. In this context, Galaxy (Afgan et al., 2018) could offer the tools we used in this study, in a more user-friendly way, not requiring the use of the command line. Additionally, Galaxy allows to compile workflows which can be shared amongst laboratories, contributing to accessibility and reproducibility. The search for unnatural associations in the proposed workflow is still manual and time-consuming. Other approaches that can be automated could be developed. It needs to be investigated to which extent these could be incorporated into a universal Galaxy workflow, suited for all GMM samples. Moreover, a more extensive analysis, e.g. including SNP-based analysis, could be included to unequivocally prove that a strain detected in the metagenomics sample is identical to a previously characterized and sequenced GMM isolate. Given the current error rate of the long read sequencing, this would be more suited for the short read sequencing only. However, it was shown that for determination of GMM genomes, the long reads help to obtain a more contiguous *de novo* assembly (Berbers et al., 2020). Rapid advances in bioinformatics tools available for ONT data (e.g. basecalling, assembly, polishing) might decrease the error rate on the long reads, which affected the target coverage observed for some reads after Flongle or MinION sequencing of a sample with lower DNA concentration (GMM16). However, this also comes with a cost as developed analysis pipelines might have to be reviewed and updated often. Hybrid assembly, thereby combining the assembly advantage of long reads with the accuracy of short reads, could ameliorate this issue. Although theoretically possible, hybrid assembly was, however, not conducted in this study as it would currently still represent a very high cost to be used routinely by enforcement laboratories. This might change in the future. Moreover, our analysis was only conducted on samples which most probably only contained one species (*B. subtilis*), and it has not yet been tested on more complex samples, in which unnatural associations might be less obvious to detect and genomes even more challenging to assemble. The detection of distinct closely related species and unnatural associations in more complex samples would require further development of appropriate analysis tools and databases. Generally, this open approach can in the future be applied to other GMM used to produce fermentation products like food enzymes. This requires that the corresponding sequence data is available in public databases, as it is able to detect any species and AMR genes / vector present in a sample based on the condition that reference data to compare with is available. Consequently, we believe that, if GMMs cannot fall under the GMO regulation, thereby resulting in no identification method being available (European Parliament and the Council of the European Union, 2003b, 2003a), sharing of information from the industry on all used vectors and species and sequences of GMMs confidentially reported to

EFSA with the enforcement laboratories and/or the competent authorities would greatly help in the development of new detection methods, including metagenomics. Indeed, this would increase the list of sequences of genes or shuttle vectors and of known GMM species to look for, thereby facilitating the open approach offered by metagenomics. Such a database would also allow investigating more closely whether specific species found in a sample using taxonomic methods are linked to misclassifications, contaminations or genetic introductions from other species. Also a database of previously sequenced GMM isolates should be constructed as this will also provide more GMM genomes to map the metagenomics reads to. In this study the reference genome most probably linked to the samples was known and therefore could be used as a final confirmation.

In conclusion, this proof of concept study delivered a novel way to detect GMMs in food/feed products using shotgun metagenomics, by uncovering unnatural associations linked to the presence of typically used AMR genes and identification of the species. This could all be achieved with the analysis of one sequencing reaction. This confirms the hypothesis of this work. Therefore, this approach would fit within the workflow used by enforcement laboratories when detection of DNA and qPCR screening led to the suspicion of the presence of an unknown GMM such as for sample GMM16, when no isolate can be obtained (i.e. no possibility to do WGS of the isolate to confirm the GMM) and a DNA walking strategy is too laborious and neither successful nor possible because the anchor is not known (Fig. 6.5). The proposed shotgun metagenomics approach allows the identification and characterization of GMMs. Theoretically, this method can replace the currently used qPCR first and second line analyses steps in the enforcements labs. This includes the detection of AMR genes or event-specific markers for which no qPCR method has been developed yet and the identification of the species, which is currently not a standard procedure. However, until the metagenomics approach is appropriately validated, currently it would rather be used by the enforcement laboratories as an orientation step, requiring confirmation of the findings by PCR and/or Sanger sequencing. With additional protocol optimization allowing longer read lengths in the future, MinION sequencing might allow the immediate detection of full-length AMR genes, thereby supporting risk assessment and a complete *de novo* assembly of the genetically modified strain. This will contribute to an open approach of generalized detection and characterization of unknown GMMs in microbial fermentation products.

Acknowledgements

We would like to thank the technicians from the Belgian enforcement laboratory (Transversal Activities in Applied Genomics, Sciensano) for the screening of the samples as well as the NGS sequencing platform from Transversal Activities in Applied Genomics (Sciensano). We also acknowledge Patrick Philipp and his team from Service Commun des Laboratoires (France) for access to the sample of GMM14 (RASFF2014.1249).

Supplementary data

Supplementary Materials S6.1-S6.5 can be found online at https://www.sciencedirect.com/science/article/pii/S2666566221000149.

CHAPTER 7
Metagenomics to detect and characterize viruses in food samples at genome level? Lessons learnt from a norovirus study

This chapter was adapted from the previously published manuscript:
Buytaers, F.E.; Verhaegen, B.; Gand, M.; D’aes, J.; Vanneste, K.; Roosens, N.H.C.; Marchal, K.; Denayer, S.; De Keersmaecker, S.C.J. De Metagenomics to Detect and Characterize Viruses in Food Samples at Genome Level? Lessons Learnt from a Norovirus Study. Foods 2022, 11, 3348, doi:10.3390/foods11213348.

Authors’ contributions:

F. E. Buytaers worked on the conceptualization, data curation, formal analysis, investigation, methodology, resources, software, visualization and writingof the original draft. B. Verhaegen helped for the conceptualization, data curation, resources and validation. M. Gand participated in the methodology and resources. J. D’aes gave insight for the investigation and software. K. Vanneste took part in the investigation, resources and software. N. H. C. Roosens participated in the funding acquisition, project administration and validation. K. Marchal was involved in the supervision. S. Denayer helped for the conceptualization and validation. S.C.J. De Keersmaecker was involved in the conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, supervision, validation and writing of the original draft. All authors have read and agreed to the published ver-sion of the manuscript.

Abstract:

In this proof of concept study on food contaminated with norovirus, we investigated the feasibility of metagenomics as a new method to obtain the whole genome sequence of the virus and perform strain level characterization but also relate to human cases in order to resolve food-borne outbreaks. We tested several preparation methods to determine if a more open sequencing approach, i.e. shotgun metagenomics, or a more targeted approach, including hybrid capture, was the most appropriate. The genetic material was sequenced using ONT with or without adaptive sampling, and the data was analyzed with an in-house bioinformatics workflow. We showed that a viral genome sequence could be obtained for phylogenetic analysis with shotgun metagenomics if the contamination load was sufficiently high or after hybrid capture for lower contamination. Relatedness to human cases goes well beyond the results obtained with the current qPCR methods. This workflow was also tested on a publicly available dataset of food spiked with norovirus and hepatitis A virus. This allowed us to prove that we could detect even less genome copies and two viruses present in a sample using shotgun metagenomics. We share the lessons learnt on the satisfactory and unsatisfactory results in an attempt to advance the field.

7.1. Introduction

Foodborne viruses, in particular noroviruses, have been described as the contaminant causing the largest number of cases of foodborne diseases worldwide (WHO, 2015). Norovirus was one of the most frequently reported causative agents for foodborne outbreaks in Europe in the past years (EFSA, 2019a, 2021c, 2021d), with frozen soft fruits and shellfish as important sources (De Keuckelaere et al., 2015; Bartsch et al., 2018). Norovirus is constituted of a positive RNA strand of about 7 thousand bases long. It is characterized based on dual typing of the ORF1 and ORF2 regions (Desdouits et al., 2020a) into genogroups and genotypes. The genogroups GI and GII, along with GIV (less common), can be pathogenic in humans, while norovirus GV is harmful in murines. The minimal infectious dose for norovirus is approximately a thousand viral particles (Bartsch et al., 2018). Viruses need host cells to replicate, making it more difficult to use culture-based detection methods for viral foodborne contaminants. Therefore, the conventional approach to detect and characterize these viruses in food is real-time polymerase chain reaction (qPCR) (ISO: International Organization for standardization, 2019). This method allows to detect specific genetic fragments of the pathogen after reverse transcription of the extracted RNA to complementary DNA (cDNA). If the result is positive, the test can be followed by a Sanger sequencing of specific regions of the genome to obtain the genotype of the virus (Mathijs et al., 2011). However, detection of norovirus in food leftovers that might be linked to an outbreak has been reported as challenging because of the low contamination dose and the heterogeneity of the contamination (Vivancos et al., 2009; Chen et al., 2016; Morgan et al., 2019), and even when the virus can be detected, it cannot always be further typed to be compared with the human cases (Baert et al., 2011). Moreover, these conventional approaches do not deliver a full characterization of the complete genome of the virus, and therefore do not allow determining accurate relatedness of cases by performing phylogeny, which is often required in outbreak investigation. New outbreaks have been found to be caused by recombinant noroviruses, for which the sequencing of the overlapping region between ORF1 and ORF2 is necessary for a correct genotyping (Desdouits et al., 2020a). Finally, Hepatitis A virus (HAV), another common foodborne virus (EFSA, 2021d), shares some similarities with norovirus but also the same contamination routes. Both RNA viruses can be tested simultaneously in a multiplex qPCR assay (Desdouits et al., 2020a)d. However, this is not done systematically, and other viruses or pathogens might be present in the food without being detected. Therefore, the conventional methods have their limitations for foodborne virus detection and characterization.

In recent years, metagenomics approaches, based on the sequencing of all genetic material of a sample, have been developed as an alternative method for various applications including the detection and characterization of viruses (Greninger et al., 2015; Rose et al., 2015; Couto et al., 2018; Lewandowski et al., 2019). The study in food samples, with a commonly reported low contamination dose compared to clinical samples, would be particularly challenging. In 2018, Bartsch et al. studied frozen strawberry samples linked to a norovirus outbreak. Out of 29 million sequencing reads, only 2 could be matched to the sequences of norovirus from patients with high identity because mostly plant and bacterial material was sequenced (Bartsch et al., 2018). Similarly, Yang et al. investigated norovirus and HAV that were artificially added at low concentrations in celery with a shot-gun metagenomics approach (Yang et al., 2017). They were able to infer the genotype of the spiked strains by mapping to a database of norovirus and HAV. However, a more open profiling approach did not succeed on their dataset and they did not obtain a genome to perform relatedness studies in case of outbreaks. To circumvent the low amount of viral reads, an-other previously documented approach is shotgun metagenomics sequencing after whole transcriptome amplification (WTA). Such a method was tested for the detection of viruses on various food matrices (Aw et al., 2016; Cibulski et al., 2021) but norovirus was not detected possibly due to an insufficient sequencing depth. Most of these studies have used short read sequencing. However, real-time long-read sequencing from Oxford Nanopore Technologies (ONT, using MinION and Flongle) might offer an interesting alternative for a lower price and turnaround time, and the possibility to have bigger fragments of the virus genome sequenced in one read. Furthermore, ONT has recently integrated a new mode called adaptive sampling to its GridION devices, allowing to selectively sequence DNA based on the similarity to reference sequences provided to the instrument (Martin et al., 2022). This might offer targeted sequencing of the pathogen(s) present in the food matrix, without targeted RNA or cDNA preparation. For this purpose, the method would need a curated database of possible foodborne pathogens (bacteria, virus, parasite…), but this application has not been tested yet. Alternatively, some studies have focused on the enrichment in viral load before sequencing. This can be conducted before the RNA extraction protocol (Conceição-Neto et al., 2015; Bal et al., 2018; Vibin et al., 2018). In particular, ultracentrifugation was performed to enrich for viral particles in the above mentioned study by Yang et al. (Yang et al., 2017). Other studies have intended at specifically increasing the viral genetic material after the RNA extraction. One of the methods that has been previously presented, is based on the removal of the background ribosomal RNA (rRNA). This showed successful on clinical samples in which the human rRNA was depleted, and even allowed to perform phylogenetics on the detected strains (Bavelaar et al., 2015; Shah et al., 2020). It was also used for the detection of plant viruses after removal of the plant rRNA using FastSelect Plant and sequencing on Flongle flow cells (Liefting et al., 2021). An alternative method is to use beads to capture polyadenylated RNA, as norovirus and HAV present a poly(A) tail. This has been described to increase the norovirus loads in stool samples 40-times (Fonager et al., 2017), but it has not yet been tested for the lower contamination dose in food samples. Finally, target enrichment of a virus of interest is possible using probes to capture the cDNA by hybridization, and washing away of the non-bound DNA. This has already been performed for norovirus in clinical but also sewage samples using probes developed specifically for human noroviruses proposed in the SureSelect products (Brown et al., 2016; Van Beek et al., 2017; Strubbia et al., 2019). This method provided a high coverage of the norovirus genome in these samples, however it has not yet been tested on food matrices. Moreover, very few studies have accompanied the hybrid capture with long reads sequencing (Eckert et al., 2016).

The goal of this proof of concept study was to test several sample preparation and sequencing methods previously described, to determine which workflow could be used to detect and characterize foodborne viruses in food samples and allow to obtain relatedness by phylogeny, taking norovirus in food matrices as a case study. Our study aims at serving as guidance for future work in the field. Therefore we tested if we could obtain the genome of the virus with each method, which had not all been used for low contamination levels in food previously. We then performed a relatedness study with a phylogenetic tree when the viral genome was obtained, which was only previously performed in very few studies for viral food contamination. We present this work with the intention of providing a workflow that could later on, after thorough validation on a large set of samples, be implemented in routine setting. Therefore, we did not alter the validated RNA extraction protocol stated in ISO 15216-2 (ISO: International Organization for standardization, 2019), currently used in the national reference laboratories (NRLs). Moreover, while most studies have been previously performed using Illumina sequencing, we decided to perform the sequencing on ONT flow cells and Flongles, because of the limited amount of samples received at the Belgian NRL for norovirus detection. A bioinformatics workflow was developed in-house in order to analyse the sequenced data obtained from different samples (different food matrices, different contamination loads) with each sample preparation or sequencing method of this proof of concept. It was also tested on a publicly available dataset of another food matrix (co-)spiked with norovirus and/or HAV at different contamination loads. The bioinformatics analysis included data quality checks and filtering, profiling to detect the presence of a virus in the sample with-out *a priori* knowledge, reference-based mapping and building of a consensus sequence. This sequence was then typed and used for phylogenetic investigation. This relatedness characterization allows going well beyond the results obtained with the current methods of analysis of norovirus in food.

7.2. Materials and Methods

7.2.1. Samples

One kilogram of frozen raspberries were bought at a local store and was divided in parts of 25 g that were used for extraction as such (bk, Figure 7.1) or spiked with 5 lenticule discs of human norovirus GI.7 (Public Health England culture collection, Salisbury, UK, mean concentration of 1.9×10^4 genome copies per lenticule disc), or with 100 µl of murine norovirus GV from the VIRSeek Murine Norovirus kit (Eurofins Genescan Technologies GmbH, Freiburg, Germany, mean concentration of 10^8 genome copies per ml). This spiking led to a concentration of 10^5 genome copies per 25 g for the norovirus GI (hunov, Figure 7.1) and 10^7 genome copies per 25 g for the norovirus GV (munov, Figure 7.1). Two biological replicates of the spiking with the human norovirus (hunov1 and 2, Figure 7.1) and the murine norovirus (munov1 and 2, Figure 7.1) were performed, on the same batch of raspberries.

A shellfish (i.e. bivalve) sample (bivalve, Figure 7.1) received at the Belgian NRL and naturally contaminated with norovirus (positive in qPCR with a Cq of 34 for the genogroup GII) was also included in the study.

7.2.2. RNA extraction

For the blank and the spiked raspberries (bk, hunov, munov), ISO 15216-2 was followed with the recommendations for soft fruits, but without addition of mengovirus as process control during our method development, to increase the chance of sequencing norovirus genetic material. This protocol consists in several steps of shaking, incubation, centrifugation and pH adjustment (ISO: International Organization for standardization, 2019). The final aqueous phase was used for RNA extraction.

For the bivalve sample, the sample preparation also followed ISO 15216-2 but for bi-valves, and the addition of mengovirus as process control was also omitted. The sample preparation consisted in addition of proteinase K to 2 g of starting material, followed by centrifugation. The supernatant was used for RNA extraction.

The RNA extraction was conducted using the Nuclisens MiniMAG kit according to the manufacturer's instructions (BioMérieux, Marcy-l'Étoile, France), which is the accredited procedure at the Belgian NRL. The final RNA of the raspberry samples was eluted in 100 µl of elution buffer. The RNA extraction of the spiked (and blank) raspberry samples was repeated 5 times and the eluates were pooled in order to have sufficient genetic material for subsequent tests (a total of 500 µl).

The presence of norovirus in the RNA pools was analyzed with qPCR (Figure 7.1, Table 7.1) using NovGI/GII @ceeramTools food kit multiplex (Biomérieux, Marcy-l'Etoile, France) to detect the human norovirus in accordance with the specifications of the ISO 15216-2, or using the VIRSeek Murine Norovirus kit (Eurofins Genescan Technologies GmbH, Freibourg, Germany) to detect the murine norovirus.

7.2.3. Genetic material preparation

Six sample preparation workflows have been tested in this study (Figure 7.1):

7.2.3.1. Method A: Poly(A) RNA capture

The polyadenylated RNA was captured from the total RNA using DynaBeads mRNA DIRECT Purification Kit (ThermoFisher Scientific, Waltham, USA) following the protocol described by Fonager et al. (Fonager et al., 2017). The captured and eluted RNA was tested for the presence of norovirus with the same qPCR as described for the total RNA, and then reverse tran-scribed using SuperScript IV (Invitrogen, Thermo Fisher Scientific, Waltham, Massachusetts, USA) with random hexamers following the manufacturer's instructions for the first strand. The second strand was then synthetized using the NEBNext Ultra non-directional RNA second strand synthesis module (New England Biolabs, Ipswich, Massachusetts, USA) following the

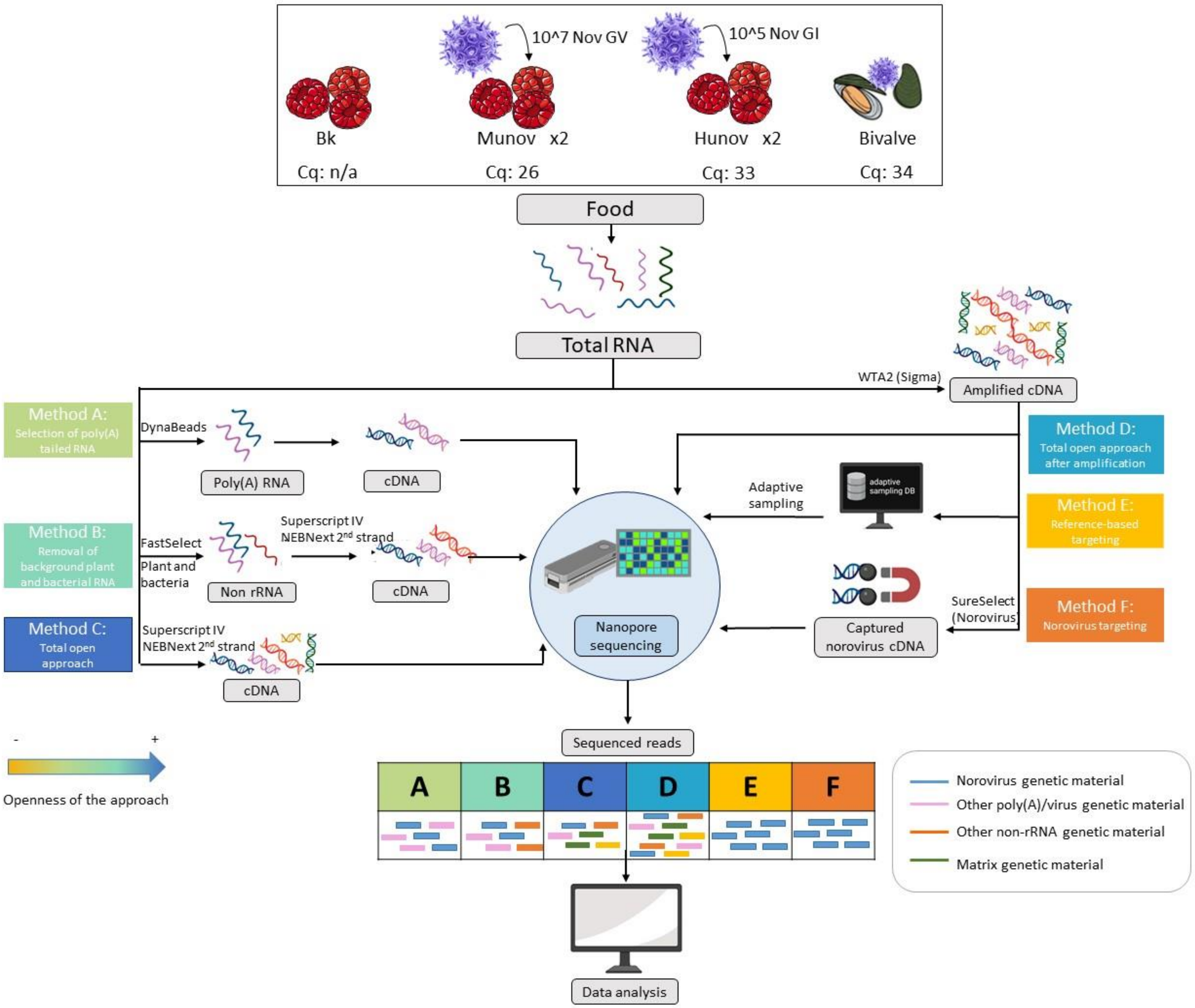

10^7 Nov GV
10^5 Nov GI
Bk
Munov x2
Hunov x2
Bivalve
Cq: n/a
Cq: 26
Cq: 33
Cq: 34
Food
Total RNA
WTA2 (Sigma)
Amplified cDNA
Method A:
Selection of poly(A) tailed RNA
DynaBeads
Poly(A) RNA
cDNA
Method B:
Removal of background plant and bacterial RNA
FastSelect
Plant and bacteria
Non rRNA
Superscript IV
NEBNext 2nd strand
cDNA
Method C:
Total open approach
Superscript IV
NEBNext 2nd strand
cDNA
Method D:
Total open approach after amplification
Method E:
Reference-based targeting
Method F:
Norovirus targeting
Adaptive sampling
adaptive sampling DB
SureSelect (Norovirus)
Captured norovirus cDNA
Nanopore sequencing
Sequenced reads
A
B
C
D
E
F
Norovirus genetic material
Other poly(A)/virus genetic material
Other non-rRNA genetic material
Matrix genetic material
-
+
Openness of the approach
Data analysis

Figure 7.1. Overview of the different samples and methods of preparation of the genetic material tested in the study, starting from the total RNA extracted from the food sample following ISO 15216-2*. Raspberries were spiked with norovirus GI at a concentration of 10^5 genome copies per 25g (hunov, Cq 33), norovirus GV at a concentration of 107 (munov, Cq 26), or kept as a blank (Bk, not detected by qPCR). The spiking was repeated twice. A naturally contaminated bivalve sample (bivalve, Cq 34) was also investigated. Each method is explained in detail in section 2 (materials and methods). Three methods are based on the total RNA (methods A-B-C) and three methods are based on the whole transcriptome amplification of the total RNA (methods D-E-F). The theoretical reads output after sequencing is displayed for each method. The methods are classified per openness of the approach with a color code (orange least open, blue most open). The blue RNA/cDNA/reads represent the genetic material from norovirus in the sample while the other colors represent genetic materials from other origins, as indicated in the figure.*

manufacturer's instructions. The prepared cDNA was cleaned using AMPure beads (Beckman Coulter, Brea, California, USA) at a ratio of 1:1 and two rounds of washing with 200 µl 70% ethanol were conducted. It was then eluted in nuclease-free water in the same volume as the starting volume after leaving at room temperature for 2 minutes.

7.2.3.2. Method B: plant and bacteria rRNA depletion

Plant and bacterial ribosomal RNA was depleted from the total RNA using the Fast-Select rRNA Plant and 5S/16S/23S kits (Qiagen, Hilden, Germany). The protocol of the FastSelect 5S/16S/23S was followed with addition of 1 µl of FastSelect Plant to the mix and no fragmentation. The reverse transcription of the two strands of cDNA was then conducted on the non-rRNA as described in method A. The prepared cDNA was cleaned using AMPure beads as for method A. The cDNA was then tested for the presence of human or murine norovirus with qPCR as described for the total RNA.

7.2.3.3. Method C: shotgun cDNA

The first-strand cDNA synthesis was performed on the pooled total RNA using SuperScript IV and the second strand of cDNA was synthetized using the NEBNext Ultra non-directional RNA second strand synthesis module as described for method A. The prepared cDNA was then cleaned using AMPure beads similarly as for method A.

7.2.3.4. Methods D and E: shotgun amplified cDNA

The pooled total RNA was reverse transcribed and amplified using the whole transcriptome amplification 2 kit (Sigma-Aldrich, Saint-Louis, Missouri, USA) according to the manufacturer's instructions. The prepared cDNA was cleaned using AMPure beads in the same way as for method A. The choice of the amplification kit has previously been shown to have an impact when sequencing long reads (Russell et al., 2018), and the WTA2 from Sigma-Aldrich was recommended over other products to avoid the sequencing of chimeric junctions that can be created during a ligation step.

7.2.3.5. Method F: Amplified norovirus captured cDNA

After amplification and beads cleaning as described in method D, the amplified cDNA was sheared to a size of about 1 kb using covaris microTUBE AFA Fiber pre-slit snap-cap 6x16mm PN 520045 with the insert microTUBE 130µl, with peak incident power 50, duty factor 2%, 200 cycles per burst for 30 seconds (Covaris, Woburn, USA). Although not ideal for long reads sequencing, the cDNA shearing was recommended for the hybridization. SureSelect XT2 was then used to capture the norovirus sheared cDNA with the PanNoro panel of probes (Agilent, Santa Clara, California, USA) according to the manufacturer's instructions including recommendations for long reads sequencing. The amplification was necessary to have sufficient starting cDNA material for the SureSelect protocol.

7.2.4. Long read sequencing

7.2.4.1. Library preparation

All samples to sequence were prepared using the Ligation sequencing kit for genomic DNA (SQK-LSK109; Oxford Nanopore Technologies, Oxford, United Kingdom) according to the manufacturer's recommendations for the specific flow cell used for sequencing. When the recommended input amount of genetic material was not met, the maximum volume of starting material (48 µl) was used.

Several sequencing methods were tested in this study: sequencing on MinION flow cells, without or with adaptive sampling (see method E), and sequencing on Flongle flow cells.

7.2.4.2. MinION sequencing

The prepared libraries (except for method E) were loaded on one Spot-ON MinION flow cell (FLO-MIN 106D; R9.4.1 version) per sample, and a 72 hours sequencing run was started on a Mk1C or a GridION device.

7.2.4.3. Method E: Adaptive sampling

The prepared libraries of shotgun amplified cDNA (prepared following the same protocol as described for method D) were loaded on one MinION flow cell (R9.4.1) per sample on a GridION device. A 72 hours sequencing run was started using the adaptive sampling option of the MinKNOW software, with a fasta file of norovirus and hepatitis A reference genomes from NCBI (listed in supplementary materials 1). The number of reads, median read length, median read quality and read length N50 for each sample are presented in Figure 7.1. All reads were analyzed as such ("all reads") or only the reads characterized as "stop_receiving" from the adaptive sampling csv file were analyzed separately. The "stop receiving" reads correspond only to reads that matched the database, and not the reads that have started to be sequenced before a decision was made by the instrument about their resemblance to the database.

7.2.4.4. Flongle sequencing

The libraries prepared with adjusted volumes for Flongle sequencing were loaded on one Flongle flow cell FLO-FLG001 (R9.4.1 version) per sample, and a 24 hours sequencing run was started.

7.2.5. ONT Data analysis

The data analysis of the ONT sequenced reads (Figure 7.2) started with basecalling using Guppy version 5.0.7 (Oxford Nanopore Technologies) in super high accuracy mode on a GPU server. Statistics were obtained from the basecalled reads using NanoPlot version 1.36.2 (De Coster et al., 2018) for quality assessment (Table 7.1). The reads were further filtered to retain only those with a quality higher than 7 using NanoFilt version 2.8.0 (De Coster et al., 2018).

The high quality reads were then assembled with Megahit version 1.1.3 (Li et al., 2015) using as k-list 21,33,55,77,99,127,155,183,211,239. This tool was selected due to the relatively low obtained read lengths that were not compatible with most assemblers designed for long reads. The k-list was designed to cover the diversity of read lengths that we obtained (see Table 7.1). Taxonomic classification was conducted on the contigs using Kraken2 version 2.1.1 (Wood et al., 2019) with default parameters and an in-house database containing all NCBI RefSeq Genome entries with the “Complete Genome” assembly level (database accessed February 11, 2021;(O’Leary et al., 2016)) accession prefixes NC, NW, AC, NG, NT, NS, and NZ of the following taxonomic groups: archaea, bacteria, fungi, human, protozoa, and viruses. To determine if a foodborne pathogen was present in the sample, manual inspection of the Kraken2 results was performed. Here, we only report the number of contigs that corresponded to norovirus. The profiling was conducted on the contigs obtained with Megahit instead of the reads in order to increase the trust in the result of the presence of species at very low con-centration, such as the low spiked concentration of the pathogenic virus of interest (Tran and Phan, 2020).

All high quality reads were then inputted in Mash version 2.2 (Ondov et al., 2016) to estimate the distance compared to the entire RefSeq database (refseq.genomes.k21s1000.msh) using mash screen with standard parameters. The result was sorted, and the best hit corresponding to the pathogen detected with Kraken2 (norovirus or hepatitis A) was used for reference-based assembly. The reads were mapped to the reference using BWA-MEM version 0.7.17 (Li, 2013) with the ont2d parameter. Variant calling was then performed using Bcftools version 1.9 (Danecek et al., 2021). Bcftools mpileup was used with the parameters –A –B –q 0 and –Q 0. Bcftools call was used to output variants sites only, with the ploidy parameter set as 1. The creation of the consensus sequence was executed with bcftools consensus using the reference previously selected with Mash. Samtools version 1.9 (Danecek et al., 2021) was used to calculate the read depth. Subsequently, the norovirus consensus sequence was typed using the online Norovirus Typing Tool version 2.0 (rivm.nl/mpf/typingtool/norovirus/ (Kroneman et al., 2011)). Finally, after multiple sequence alignment using ClustalW on MEGA-X (Kumar et al., 2018) with the default parameters, a maximum likelihood phylogenetic tree was constructed based on the consensus sequences as well as norovirus sequences from NCBI (Sayers et al., 2022) on MEGA-X with 100 bootstraps, using the Tamura-Nei model with partial deletion and other default parameters.

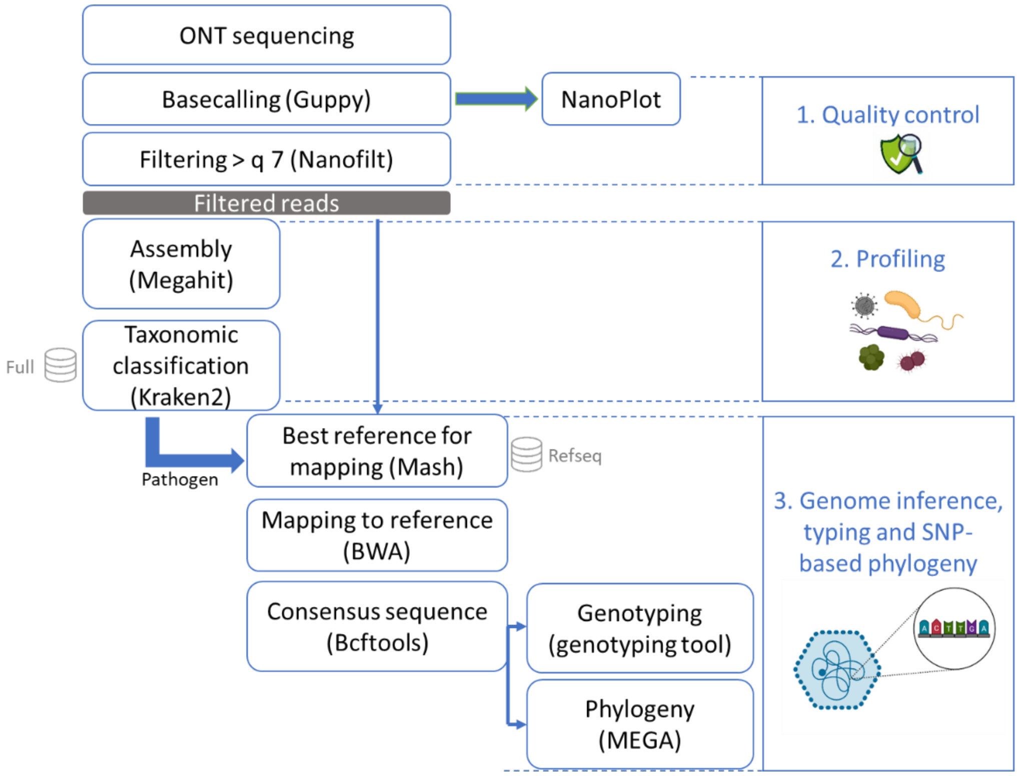

ONT sequencing
Basecalling (Guppy)
NanoPlot
Filtering > q 7 (Nanofilt)
Filtered reads
1. Quality control
Assembly (Megahit)
Full
Taxonomic classification (Kraken2)
2. Profiling
Pathogen
Best reference for mapping (Mash)
Refseq
Mapping to reference (BWA)
Consensus sequence (Bcftools)
Genotyping (genotyping tool)
Phylogeny (MEGA)
3. Genome inference, typing and SNP-based phylogeny

Figure 7.2. Bioinformatics workflow after sequencing with ONT (MinION or Flongle). *After sequencing, the data is basecalled and checked for quality control. The reads are then all assembled and the contigs are put through a taxonomic classification tool with a database of mammals, archaea, bacteria, fungi, human, protozoa, and viruses. Once a pathogen is recognized in this pro-filing step, all reads are mapped to the RefSeq database using Mash to obtain the best matching reference genome. This reference genome is used to perform mapping and to build a consensus sequence which can then be typed and used for phylogenetic analysis.*

7.2.6. Adaptation of the bioinformatics workflow to the analysis of Illumina data and the detection and characterization of HAV

In order to test our method on samples in which two viruses were present, but also with a spiking at a lower concentration, we used the publicly available data from 13 samples sequenced on Illumina MiSeq from Yang et al. (Yang et al., 2017), downloaded from the NCBI Sequence Read archive under BioProject PRJNA377525. Briefly, Yang et al. spiked celery with norovirus GII at various concentrations (i.e. 10^3 to 10^5 viral RNA copies in 50 g) and co-spiked two strains of norovirus GII (GII.4 and GII.6) or one strain of norovirus GII and one strain of hepatitis A virus (HAV HM175/18f.1, genotype IB). The RNA was extracted using QIAamp Viral RNA kit (Qiagen, Hilden, Germany) after an ultracentrifugation-based viral particle enrichment. The obtained RNA was reverse transcribed and sequenced on an Illumina MiSeq generating paired-end 100 bp reads (Yang et al., 2017). It was analyzed with the same data analysis workflow, with some modifications due to the difference in sequencing technology and the reads being paired: Basecalling and quality filtering were replaced by trimming using Trimmomatic version 0.38 (Bolger et al., 2014) on paired-end reads using truSeq3 adapter sequence, the assembly was performed with megahit with the conventional k-list (21,29,39,59,79,99,119,141), and BWA mem was used without the ont2d parameter. The online Hepatits A Virus Genotyping tool version 1.0 (https://www.rivm.nl/mpf/typingtool/hav/ (Kroneman et al., 2011)) was used to type the consensus sequences obtained for the hepatitis A virus in the samples.

7.3. Results

In this study, we have spiked raspberries with two genogroups of norovirus (norovirus GI at 10^5 genome copies in 25 g of raspberries, hunov, and norovirus GV at 107 genome copies in 25g of raspberries, munov) to represent two contamination levels, of qPCR Cqs of respectively 33 and 26. The blank raspberries were also investigated as well as a bi-valve sample naturally contaminated with norovirus GII with a qPCR Cq of 34 (lower contamination level than the spiked samples). Aiming specifically to present a method that could be later applicable in routine settings, we have decided to follow the RNA extraction method used at the Belgian NRL, covered in ISO 15216-2. As we did not alter the RNA ex-traction method, we tested several post-extraction methods for genetic material preparation (Figure 7.1) in order to increase the presence of the virus of interest: i.e. poly(A) RNA capture (method A), plant and bacterial rRNA removal (method B), norovirus cDNA hybrid capture (method F) and whole transcriptome amplification (method D). A shotgun metagenomics protocol without specific treatment of the extract (method C) was also evaluated on the same spiked samples. All these prepared genetic materials were then sequenced using the ONT platform. Adaptive sampling during nanopore sequencing was also assessed for this case study as an alternative to the targeting of the virus with laboratory methods (method E). The main differences in input and output and how much they target norovirus are presented in Figure 7.1.

7.3.1. Selection at RNA level: subset of poly(A) RNA (method A) or non-rRNA (method B)

In order to decrease the complexity of the sample, a selection at RNA level was attempted with two different methods: capture of polyadenylated RNA (method A, Figure 7.1) and depletion of ribosomal RNA from plants and bacteria (method B, Figure 7.1). Both methods were tested on raspberry samples spiked at different contamination levels using human or murine norovirus. The resulting RNA was then tested via qPCR before reverse transcription. However, for both methods, norovirus was not detected anymore by qPCR (data not shown) while it was detected in the total RNA prior to the sample treatment. These samples were therefore not used further for sequencing.

7.3.2. An open approach: shotgun sequencing of cDNA (method C)

As the most straightforward and open approach, we tested the shotgun metagenomics sequencing of the samples after reverse transcription of the RNA (method C, Figure 7.1). This was performed on biological replicates of munov and hunov (Figure 7.1). After sequencing, tens of thousands to over a million reads were produced for the spiked samples (table 7.1) with a median read length of a few hundred base pairs and a quality of 9 to 10.

The sequenced reads were then analysed through the developed bioinformatics workflow (Figure 7.2). As our previously published workflow (Buytaers et al., 2021b) for the strain-level analysis of long-reads sequencing data was not successful on the case study of this viral contamination (data not shown), we designed an alternative data analysis workflow (Figure 7.2). After assembly of the reads, a taxonomic classification was performed on the contigs. Norovirus was detected in contigs of the cDNA sequencing of the two munov samples (Table 7.2) and in one of the samples spiked with human norovirus (hunov, Table 7.2). The sequencing of the blank (non-spiked) raspberries lead to the lowest amount of reads (bk, Table 7.1), of which none could be related to norovirus or another foodborne virus (Table 7.2). The high amounts of unclassified contigs probably corresponded to the raspberry genetic material, of which no reference was present in the database used for profiling.

The closest norovirus reference was then estimated from all reads, and this reference was used for mapping and generation of a consensus sequence, that was then typed (Table 7.2).

For munov, NC_008311.1 was the most similar reference in the RefSeq database. Thirty-five and 155 reads mapped to this reference for the first and second replicate of the spiking, covering 85 and 99% of the norovirus genome, respectively (Table 7.2). Consensus sequences of 6,304 and 7,280 bases were obtained and could be correctly typed as norovirus GV (murine norovirus).

For the first replicate of hunov, NC_031324.1 was determined as the closest reference. Eighty-six reads covered 16 of the reference genome, and a consensus sequence of 1221 bases was obtained and correctly characterized as norovirus GI. The characterization went further to describe the strain as a norovirus GI.3P3, although the lenticule that was used to spike the sample was notified as norovirus GI.7 by the supplier.

Table 7.1. Sequencing statistics for each of the sample preparation methods tested.

		Number of Sequenced Reads	Median Read Length	Median Read Quality	Read Length N50
shotgun cDNA Method C	munov1	27,406	471	9.8	634
	munov2	237,26	388	10.1	501
	hunov1	1,119,816	601	9.9	814
	hunov2	106,195	387	10.2	517
	Bk	10,143	900	9.4	1251
shotgun WTA method D	munov1	13,992,038	408	12.8	483
	munov2	2,576,015	345	12.8	421
	hunov1	15,151,030	405	13.2	609
	hunov2	22,591,785	401	12.6	542
	Bivalve	14,219,052	331	13	401
	Bk	3,898,276	310	13.2	351
WTA hybrid capture Method F	munov1	38,582,756	394	12.9	409
	munov2	28,497,819	431	18.2	494
	hunov1	12,473,896	328	12.8	338
	hunov2	14,357,765	391	12.8	407
	Bivalve	8,263,607	412	12.4	442
WTA Adaptive sampling method E	munov2_all_reads	1,535,739	333	12	368
	munov2_stop_receiving	161	453	12.6	505
	hunov2_all reads	3,269,856	370	12.9	404
	hunov2_stop_receiving	0	n/a	n/a	n/a
Flongle	munov1 wta (method D)	4,810	476	7.6	573
	munov1 wta SS (method E)	12	294	3.3	4917
	hunov1 cDNA (method C)	6846	463	7.5	527
	hunov1 wta (method D)	10,119	363	7.6	416

munov= artificial spike of raspberries with murine norovirus at 107 genome copies per 25 g (genogroup GV, RNA qPCR Cq: 26). hunov= artificial spike of raspberries with human norovirus GI at 10^5 genome copies per 25 g (RNA qPCR Cq: 33). 1 and 2: biological replicates of the spiking. Bivalve: naturally contaminated bivalve sample qPCR-positive (RNA Cq: 34) for the presence of norovirus GII. Shotgun cDNA: reverse transcription of the extracted RNA (Method C). Shotgun WTA: cDNA after whole transcriptome amplification (Method D). WTA hybrid capture: cDNA after whole transcriptome amplification with target enrichment for norovirus using SureSelect (Method F). WTA Adaptive sampling: cDNA after whole transcriptome amplification, sequenced with adaptive sampling (Method E). All reads: all reads sequenced during adaptive sampling. Stop-receiving: only reads matching to the database of norovirus and hepatitis A virus reference genomes during adaptive sampling. The results for methods A and B were not presented as no sequencing was conducted after negative qPCR result. n/a: not applicable.

Table 7.2. Results of data analysis of samples of raspberries spiked with murine norovirus (munov) in biological duplicates (munov1 and munov2), human norovirus (hunov) in biological duplicates (hunov1 and hunov2), a blank of raspberries (bk) and a naturally contaminated sample of bivalve positive for norovirus in qPCR (bivalve). *Shotgun cDNA: complementary DNA after reverse transcription (method C). Shot-gun WTA: cDNA after whole transcriptome amplification (Method D). WTA hybrid capture: cDNA after whole transcriptome amplification with target enrichment for norovirus using SureSelect (method F). WTA Adaptive sampling: cDNA after whole transcriptome amplification, sequenced with adaptive sampling (Method E). All reads: all reads sequenced during adaptive sampling. Stop receiving: only reads matching to the database of norovirus and hepatitis A virus reference genomes during adaptive sampling. n/a: not applicable because norovirus not detected in the profiling step.*

		Assembly				**Data analysis**								
		Number of Contigs	**Min Length**	**Max Length**	**N50**	**Kraken Unclassified %**	**Kraken Norovirus Contigs**	**Mash Best Norovirus Hit**	**Identity to the Reference (%)**	**Mash Matching Hashes**	**Coverage of the Reference (%)**	**BWA Number of Reads Mapping**	**Length Consensus Sequence**	**Genotype Consensus Sequence**
Shotgun cDNA Method C	**munov1**	1098	256	3409	522	58	7	NC_008311.1	88	62/1000	85	35	6304	GV
	munov2	11,220	240	8,430	484	37	8	NC_008311.1	90	103/1000	99	155	7,280	GV
	hunov1	248,221	240	13,146	553	90	1	NC_031324.1	80	8/1000	16	86	1221	GI.3P3
	hunov2	9,750	241	12,527	578	81	0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	Bk	4088	269	7492	719	83	0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Shotgun WTA method D	**munov1**	55,942	240	3494	486	84	8	NC_008311.1	90	111/1000	91	743	6694	GV
	munov2	360,178	240	2576	421	50	11	NC_008311.1	91	147/1000	98	1312	7271	GV
	hunov1	458,249	240	4737	503	93	0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	hunov2	1,023,060	240	3692	443	89	0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	Bivalve	415,513	240	4842	437	93	0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	Bk	250,101	240	4482	443	94	0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
WTA hybrid capture Method F	**munov1**	513,411	240	1681	378	84	6	NC_008311.1	91	135/1000	98	2786	7255	GV
	munov2	853,353	240	3448	372	36	16	NC_008311.1	90	183/1000	100	18,030	7381	GV
	hunov1	144,897	240	4682	349	91	2	NC_031324.1	85	32/100	40	301	3115	GI
	hunov2	328,106	240	1671	370	70	18	NC_031324.1	89	80/1000	78	2636	6071	GI
	Bivalve	149,962	240	1614	377	63	0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
WTA adaptive sampling method E	**munov2_all_reads**	210,832	240	2538	403	54	7	NC_008311.1	91	124/1000	95	711	6978	GV
	munov2_stop_receiving	15	299	1006	498	33	9	NC_008311.1	90	104/1000	61	161	4474	GV
	hunov2_all reads	358,200	240	2148	443	92	0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	hunov2_stop_receiving	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Flongle	**munov1 wta**	849	300	1452	541	89	0	n/a	n/a	n/a				
	hunov1 cDNA	1383	269	2257	538	82	0	n/a	n/a	n/a				
	hunov1 wta	1708	300	1,680	433	92	0	n/a	n/a	n/a				

The consensus sequences obtained for the two munov and hunov replicates with the different sample preparation methods were placed in a phylogenetic tree, and clustered together respectively, while separating from other norovirus GV (murine) or norovirus GI (human) reference genomes (Figure 7.3).

7.3.3. An open approach with increased input genetic material: shotgun sequencing of amplified cDNA (method D)

In order to increase the overall sequencing output, including the virus sequenced reads, the RNA was reverse transcribed with an amplification kit. All the amplified cDNA was then sequenced. This was done on the two biological replicates of munov and hunov, but also on the bivalve sample. The blank raspberries (bk) were also sequenced after the same amplification.

After amplification, the number of reads sequenced respectively for each sample was increased by 10 to 100 times (Table 7.1). The read quality increased as well, but the read length and N50 slightly decreased.

After assembly, norovirus could only be detected in the contigs of the munov samples, not in hunov, bk nor the bivalve sample (Table 7.2). NC_008311.1 was again identified as the most similar reference in the database (Table 7.2). Seven hundred fourty-three and 1312 reads mapped to this reference, covering 91 and 98% of the genome for the first and second spiking experiments. Consensus sequences of 6694 and 7271 bases were obtained and correctly characterized as norovirus GV (murine norovirus).

The consensus sequences of the two norovirus strains from the munov samples could be placed correctly in a phylogenetic tree, clustering with other murine norovirus (GV) strains but separated from other non-GV norovirus genomes (Figure 7.3).

7.3.4. Capturing norovirus in the amplified genetic material: sequencing amplified targeted cDNA (method F)

In order to increase the viral load in the cDNA, the norovirus genome can be targeted with a hybridization and capture method (SureSelect) based on a panel of probes designed based on human strains of norovirus. In order to have sufficient starting material for this protocol, the extracted RNA was first amplified like in method D. This was performed on the two biological replicates of munov and hunov, but also on the bivalve sample (Figure 7.1). This double amplification (i.e. whole transcriptome amplification and capture of norovirus cDNA) lead to the highest number of sequenced reads, with a quality above 12 but a short length of approximatively 400 bp (Table 7.1), while the shearing performed for this protocol is expected to create fragments of 1 kb.

After assembly of the reads, norovirus was detected in the contigs of munov1 and 2 and hunov1 and 2, but not in the bivalve sample.

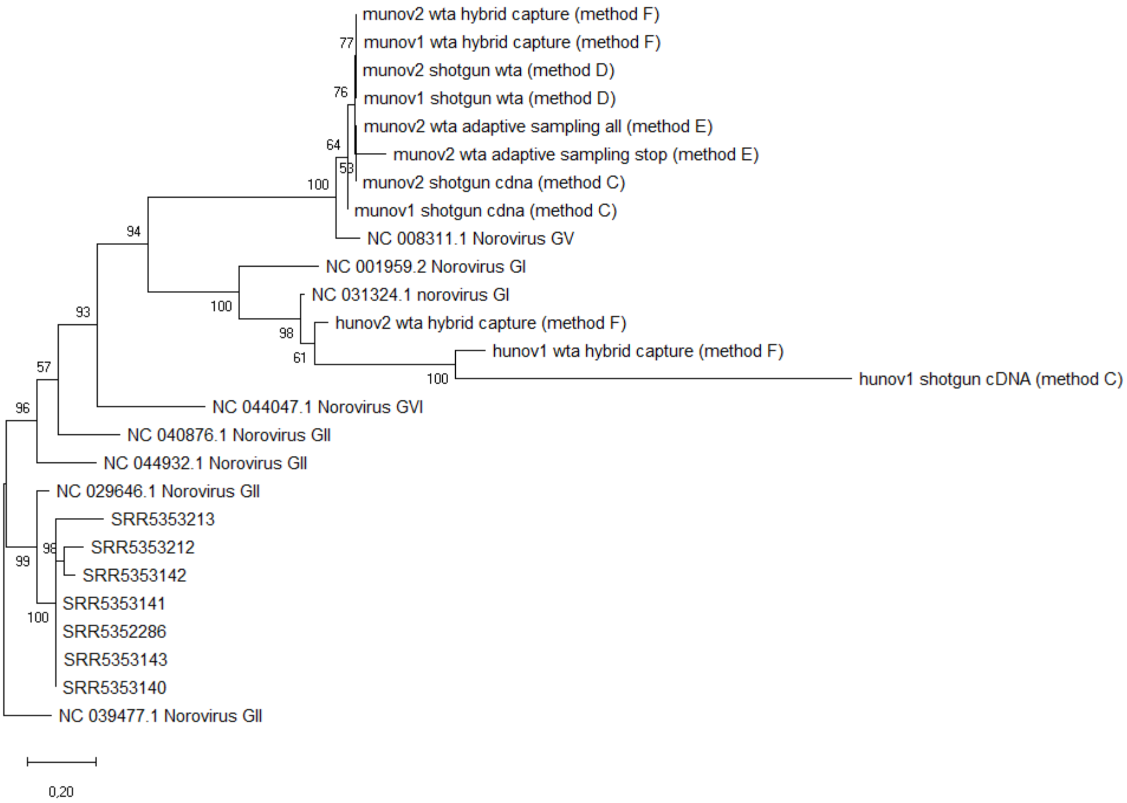

munov2 wta hybrid capture (method F)
77
munov1 wta hybrid capture (method F)
munov2 shotgun wta (method D)
76
munov1 shotgun wta (method D)
munov2 wta adaptive sampling all (method E)
64
munov2 wta adaptive sampling stop (method E)
53
100
munov2 shotgun cdna (method C)
munov1 shotgun cdna (method C)
94
NC 008311.1 Norovirus GV
NC 001959.2 Norovirus GI
93
100
NC 031324.1 norovirus GI
98
hunov2 wta hybrid capture (method F)
61
hunov1 wta hybrid capture (method F)
57
100
hunov1 shotgun cDNA (method C)
96
NC 044047.1 Norovirus GVI
NC 040876.1 Norovirus GII
NC 044932.1 Norovirus GII
NC 029646.1 Norovirus GII
SRR5353213
98
SRR5353212
99
SRR5353142
SRR5353141
100
SRR5352286
SRR5353143
SRR5353140
NC 039477.1 Norovirus GII
0,20

Figure 7.3. Phylogenetic tree of consensus sequences of all norovirus strains from metagenomics datasets and several reference strains as background. *SRR535XXX: norovirus strains obtained from metagenomics samples of spiked celery. Munov: in biological duplicates (1,2), norovirus strain obtained from raspberries spiked with murine norovirus GV. Hunov: in biological duplicates (1,2), norovirus strain obtained from raspberries spiked with human norovirus GI. Shotgun cDNA: complementary DNA after reverse transcription (method C). Shotgun wta: cDNA after whole transcriptome amplification (Method D). Wta hybrid capture: cDNA after whole transcriptome amplification with target enrichment for norovirus using SureSelect (Method F). Wta adaptive sampling: cDNA after whole transcriptome amplification, sequenced with adaptive sampling (Method E). All: all reads sequenced. Stop: only "stop receiving" reads after adaptive sampling of whole transcriptome amplification cDNA. Background strains obtained from NCBI. The scale represents the distance of 20% genetic variation.*

The closest reference to the munov samples was NC_008311.1 as obtained with the other methods (Table 7.2). Ninety-eight and 100% of its genome was covered, with 2786 and 18030 reads mapping to the reference. A consensus sequence of 7255 and 7381 bases was obtained and correctly characterized as norovirus GV (murine norovirus).

For the hunov samples, NC_031324.1 was the closest reference. Fourty and 78% of the genome was covered, with 301 and 2636 reads. A consensus sequence of 3115 and 6071 bases was obtained, and it was correctly typed as norovirus GI.

The obtained consensus sequences were placed in a phylogenetic tree (Figure 7.3). The two norovirus strains from the munov samples clustered with the other murine norovirus strains while the two norovirus strains from the hunov samples clustered with the other human norovirus strains. Both clusters separated from genomes belonging to other genogroups.

7.3.5. Selection of viral genetic material during the sequencing through the pore: adaptive sampling of the amplified cDNA (method E)

As an alternative method to the capture of the virus with the SureSelect kit (method F), the DNA fragments corresponding to the norovirus were selected during the sequencing using adaptive sampling. This method compares the read being sequenced in real-time to a database (in this case a database of norovirus and hepatitis A virus sequences, supplementary materials 1). If the read differs from the sequences in the database, the pore releases the cDNA strand and captures another cDNA to sequence. This way, the targeted species are preferentially sequenced and should be represented in higher proportions in the reads. This was performed on the second biological replicate of hunov and munov.

The complete set of sequenced reads was analyzed, but we also analyzed separately the reads characterized as “stop receiving”, which only represent the reads that matched to the reference genomes in the database. One hundred sixty-one reads were tagged as “stop receiving” for munov2, while none were tagged as such for hunov2 (Table 7.2). After assembly, norovirus could be detected by taxonomic classification only in the munov2 sample for all the reads or the stop receiving reads (Table 7.2). Mash determined NC_008311.1 as the closest norovirus reference, and 91% of its genome was covered by all the sequenced reads (compared to 98% for the same amplified DNA sample without adaptive sampling). A consensus sequence of 6978 bases could be constructed based on the 711 reads that mapped to the reference (compared to 1312 for the same sample with-out adaptive sampling), and it was typed as norovirus GV. For the stop receiving reads, only 61% of the reference genome could be covered and all the 161 reads mapped to the reference. A consensus sequence of 4474 bases was constructed and was characterized as norovirus GV.

The consensus sequence of murine norovirus obtained from all reads or just the “stop receiving reads”, could be placed in a phylogenetic tree, and clustered with other murine noroviruses from the study, separated from another norovirus GV and from noroviruses from other genogroups (Figure 7.3).

7.3.6. Flongle sequencing as a low-output less expensive sequencing alternative

Several samples were also sequenced on Flongle flow cells, in order to verify at which level of contamination a low-output less expensive sequencing alternative would be able to detect and characterize the norovirus in the samples. The number of reads sequenced on Flongles was lower than expected (a Flongle should have about 10% of pores compared to a normal flow cell and therefore we expect 10% of output) (Table 7.1). In particular, the sample of amplified and captured cDNA (method F) only presented 12 reads (Table 7.1). For that sample, no further analysis was performed. The three other samples presented a few thousand sequenced reads (Table 7.1). After assembly, no contig could be recognized as norovirus by taxonomic classification (Table 7.2).

7.3.7. Assessing our bioinformatics workflow for the analysis of a dataset containing a co-spike of norovirus and hepatitis A virus

In order to test the performance of our bioinformatics pipeline as an ‘open approach’ method, we analysed a publicly available dataset containing co-spikes of norovirus and hepatitis A virus. In 2017, Yang et al. spiked celery with norovirus GII at various concentrations and co-spiked two strains of norovirus GII or one strain of norovirus GII and one strain of hepatitis A virus. The RNA extracted from these samples was reverse transcribed and sequenced on an Illumina MiSeq (Yang et al., 2017). We conducted the data analysis on their data with the workflow we developed, adapted to the investigation of reads from Illumina.

Our results (Table 7.3) show that norovirus was detected after taxonomic classification in 11/13 samples spiked with norovirus. Hepatitis A virus was detected in 4/4 samples spiked with hepatitis A virus (co-spiked with norovirus GII).

Mash picked the closest norovirus reference as NC_029646.1 in the 11/13 samples for which norovirus was previously detected in the profiling step. A consensus sequence was obtained for 7 samples, covering 23 to 56% of the reference genome, and correctly typed as norovirus GII or GII.P4. The consensus sequence was not obtained for the 2 samples co-spiked with 2 viral species for which there was no norovirus hit with mash (SRR5353214 and SRR5353215), and for the 4 samples co-spiked with two strains of norovirus GII, for which only 1 hit was obtained with mash and the two strains could not be resolved (SRR5353144, SRR5353145, SRR5353158, SRR5353159). Norovirus was detected in the 2/4 samples co-spiked with the two viral species spiked at a higher concentration of norovirus (10^6 genome copies in 50 g of celery). The two norovirus consensus sequences obtained covered 23 and 41% of the reference genome and were correctly characterized as norovirus GI.P4. At the time of their study, Yang et al. were able to detect the norovirus GII.4 in all 4 samples co-spiked with HAV and norovirus, but they were not following a fully open approach as they were using a curated database of norovirus and HAV.

Table 7.3. Results of data analysis of samples of celery spiked with norovirus (NOV) and/or HAV (Yang et al., 2017)***.*** *-1 : first biological replicate. -2: second biological replicate. When two viruses were spiked, the best mash hit of each species was presented along with the result of the mapping and typing for each strain. n/a: not applicable (analysis was not continued because only 1 strain detected when 2 strains were spiked). Sequence read lengths: 35-100 bp.*

			Assembly				**Taxonomic classification**			**Data analysis**						
Accession Number	**Sample Description (Spike Description)**	**Number of Sequenced Reads**	**Number of Contigs**	**Min Length**	**Max Length**	**N50**	**Kraken Unclassified %**	**Number of Contigs Norovirus**	**Number of Contigs HAV**	**Mash Best Norovirus Hit**	**Identity to the Reference (%)**	**Mash Matching Hashes**	**Coverage of the Reference (%)**	**BWA Number of Reads Mapping**	**Length Consensus Sequence**	**Genotype Consensus Sequence**
SRR5352286	NOV 10^5	2,090,232	120	202	7680	783	24	1	0	NC_029646.1 (NOV)	86	40/1000	55	11,104	4137	GII
SRR5353140	NOV 10^4 -1	2,056,188	163	201	4233	611	40	2	0	NC_029646.1 (NOV)	86	43/1000	55	15,587	4142	GII
SRR5353141	NOV ^{104} -2	2,478,355	113	200	7578	898	4	1	0	NC_029646.1 (NOV)	85	37/1000	56	16,649	4209	GII.P4
SRR5353142	NOV 10^3 -1	1,879,486	1296	200	4546	509	85	3	0	NC_029646.1 (NOV)	83	19/1000	30	1416	2,23	GII
SRR5353143	NOV 10^3 -2	2,072,984	1903	200	3576	503	88	2	0	NC_029646.1 (NOV)	84	26/1000	35	1666	2618	GII.P4
SRR5353144	NOV GII.4 10^6 + NOV GII.6 10^1 -1	2,335,180	14,741	200	7444	692	94	3	0	NC_029646.1 (NOV)	84	25/1000	n/a	n/a	n/a	n/a
SRR5353145	NOV GII.4 10^6 + NOV GII.6 10^1 -2	2,546,316	13,276	200	7176	642	94	3	0	NC_029646.1 (NOV)	83	19/1000	n/a	n/a	n/a	n/a
	NOV GII.4 10^3 + NOV GII.6 10^4 -1	2,705,565	15,329	200	7076	711	94	4	0	NC_029646.1 (NOV)	72	1/1000	n/a	n/a	n/a	n/a
SRR5353159	NOV GII.4 10^3 + NOV GII.6 10^4 -2	2,203,937	13,245	200	7076	665	94	7	0	NC_029646.1 (NOV)	72	1/1000	n/a	n/a	n/a	n/a
SRR5353212	NOV 10^6 + HAV 10^6 -1	2,109,188	9122	201	7471	605	94	4	1	NC_029646.1 (NOV) NC_001489.1 (HAV)	84 99	29/1000 898/1000	41 99	3554 6524	3069 7462	GII.P4 HAV I.B
SRR5353213	NOV 10^6 + HAV 10^6 -2	1,736,984	7706	201	7471	606	93	1	1	NC_029646.1 (NOV) NC_001489.1 (HAV)	82 99	14/1000 896/1000	23 99	1713 5484	1708 7460	GII.P4 HAV I.B
SRR5353214	NOV 10^4 + HAV 10^7 -1	221,046	674	201	5684	442	89	0	2	NC_001489.1 (HAV)	99	889/1000	99	3383	7438	HAV I.B
SRR5353215	NOV 10^4 + HAV 10^7 -2	797,912	4416	201	5803	538	93	0	2	NC_001489.1 (HAV)	99	894/1000	100	5614	7468	HAV I.B

For hepatitis A virus, NC_001489.1 was the closest reference in our database for the four samples, and a consensus sequence covering more than 99% of the reference was obtained. It was correctly characterized as HAV I.B in all cases.

All obtained consensus sequences were then placed in a phylogenetic tree. All norovirus strains obtained from the samples of Yang et al. clustered together, while separating from other norovirus GII genomes (Figure 7.3). This was the expected result, as consensus sequences were obtained from samples all spiked with the same strain of norovirus GII.4.

7.4. Discussion

Foodborne viruses, and in particular norovirus, represent a major worldwide burden on our food safety. However, due to their very low contamination dose in food, they are particularly hard to detect in food products suspected to cause a foodborne outbreak. Moreover, the current methods to detect norovirus in food samples do not give the full information about its genome, nor allow relatedness analysis. In this study, we investigated metagenomics as a new alternative approach that would allow to obtain the genome of the viral pathogen present in the food sample and perform relatedness analysis with phylogenetic trees. For all samples, we performed RNA extraction according to the ISO norm currently in practice at the Belgian NRL, in order to present a protocol that could be easy to implement as an alternative for these laboratories after formal validation of the full work-flow. Because metagenomics enables sequencing all genetic material in the sample, only a few reads might belong to the virus of interest. We therefore tested different sample preparation and sequencing approaches with various degrees of targeting of the virus. We decided to sequence with the MinION or Flongle flowcells from ONT because they offer fast results (real-time sequencing) compared to e.g. Illumina sequencing, but also because they are more cost-effective when only a few samples have to be sequenced at a time (Buytaers et al., 2021b). This would all help in a further application of the protocol in routine. In order to compare the results obtained for each method, we developed a bioinformatics workflow to analyze the data without *a priori* knowledge, profiling for the pathogen in the sample, obtaining its genome, characterizing it and relating it to other sequences. This was done as a proof of concept to deliver the most suited protocol to investigate further for future implementation.

As the most open approach, we tried shotgun sequencing on the extracted RNA. We compared the results obtained with either reverse transcription (method C) or with whole transcriptome amplification (reverse transcription followed by random amplification, method D) in order to improve the input RNA amount. Amplification has been described in several studies as a necessary step for the detection of viruses with shotgun meta-genomics (Conceição-Neto et al., 2015; Li et al., 2020), and ONT sequencing in particular requires high levels of starting genetic material. In the current study, amplification of the genetic material enhanced results for the detection and characterization of norovirus in the sample spiked with murine norovirus (GV) at a concentration of 10^7 (longer consensus sequence and 10-fold increase in number of reads mapping to the reference) but not in the samples spiked at a lower con-centration with norovirus GI and in the naturally contaminated bivalve sample.

Therefore, although improving the result at high contamination level, unspecific amplification did not allow strain-level characterization and phylogeny at a lower contamination. Both with and without amplification, when the virus was detected during the profiling step, a consensus sequence could be obtained and it was typed and correctly placed in a phylogenetic tree. In one sample spiked with the human norovirus, the typing to genotype level was incorrect (based on the information we received from the supplier of the norovirus reference material), however the genogroup was correctly determined for all samples. Moreover, phylogeny allows a relatedness at a higher discriminatory level than genotyping.

In order to increase the amount of sequenced reads corresponding to the pathogen and possibly decrease the host DNA sequenced (the food), we tested more targeted approaches. Because we wanted our workflow to be applicable in a routine laboratory, we decided to keep the conventional RNA extraction workflow described in ISO 15216-2. We then tested two post-RNA extraction methods to increase the norovirus load in the samples to be sequenced. We first tested an approach that could capture the polyadenylated RNA (method A, norovirus RNA harboring a poly(A) tail) and an approach depleting the ribosomal RNA from plants and bacteria (method B). Unfortunately, these two methods did not give the expected result, as norovirus could not be detected for any of our samples within the detection limit of the qPCR after following these two protocols. The explanation for this lack of success was probably the very low contamination load of norovirus in our samples. For the first method, the few RNA fragments belonging to the virus were probably lost during the poly(A) capture and washing step. A previous study had reported very good results with this method, but it was conducted on stool samples, with a much higher dose of the virus (Fonager et al., 2017). In the case of the ribosomal RNA depletion, the cause of our lack of success was possibly also a loss of norovirus RNA due to dilution during the protocol or washing out. Moreover, we observed that raspberry was not part of the sequences used to design this plant RNA depletion kit. Indeed, although this kit gave good results in other studies, it is not universal for all plants. In fact, a study of plant viruses using the same method followed by Flongle sequencing reported lower results when analyzing strawberries than peas (Liefting et al., 2021). Moreover, FastSelect does not exist for other eukaryotes (like bivalves) and therefore this method could not be applicable in a routine laboratory setting handling various types of food matrices.

Aiming to further improve the results, two other methods were tested that targeted directly the virus of interest, at the cDNA level, after amplification: a target enrichment using SureSelect based on capture using a panel of probes designed for human norovirus (method F), and an adaptive sampling during the nanopore sequencing based on a database containing references of norovirus (method E). The amplification was necessary for both methods in order to have sufficient input genetic material for the protocols. Moreover, although a protocol adapted for ONT sequencing was available for the SureSelect (received from the R&D team of Agilent), this method primarily aims at preparing a library for short reads sequencing and the cDNA had to be fragmented to an average size of 1 kb, which is not ideal for subsequent long reads sequencing. Our results showed that this double amplification was the only method able to detect and characterize the norovirus at both levels of artificial contamination (10^5 and 10^7 genome copies per 25g of fruit). The obtained consensus sequence of the virus could be typed and correctly placed in a phylogenetic tree. This sample preparation method was, however, not able to lead to the detection norovirus at a lower concentration in the naturally contaminated bivalve sample. SureSelect was indeed previously shown to work better with genome copy inputs higher than 10^4 in previous tests from the company (Williams et al., 2019). Although this is unfortunate as the contamination load in food samples can be very low, this is in agreement with the results we obtained, as we could not detect norovirus after SureSelect enrichment in the sample with the lowest contamination (the bivalve sample). Yet, this approach is very selective as it can only target norovirus and no other viral pathogen in the sample, and it is based on a panel of probes, which might not recognize a novel variant. For these reasons, a new method associated with ONT sequencing was tested: adaptive sampling. We tested this approach with a database of noroviruses and hepatitis A viruses in order to be more open than the SureSelect approach targeting only noroviruses. Ideally, for our application, a database of references of all food pathogens should be provided to the software. In our case, we could see that this method did not improve the results compared to those obtained on the same genetic material without adaptive sampling. This is probably explained by the shortness of our cDNA fragments, as at least 400 bp have to pass through the pore for the software to determine if the DNA strand resembles the reference(s) (Marquet et al., 2022), but our mean read length was close to 400 bp. For the sample spiked with human norovirus, no read was recognized as norovirus while for the sample spiked with murine norovirus, 161 reads were tagged as “stop receiving” during adaptive sampling which means they corresponded to a reference in our database and the sequencing continued for this DNA fragment. This still represented a loss compared to the 711 reads that mapped to the reference when using all reads sequenced in the run. This could be improved by producing longer cDNA fragments to sequence or if the number of bases necessary for the tool to make its decision decreases in further updates from ONT. Consequently, unfortunately, the adaptive sampling was not a usable alternative for this case study at the time the experiments were conducted.

As our method was able to obtain results after MinION sequencing of several samples, the Flongle was tested as a less expensive alternative sequencing approach. Although we had an acceptable amount of active pores for Flongle sequencing, very few reads were obtained compared to the MinION sequencing (less than the 1/10th that would be expected from the difference in number of pores), and norovirus could not be detected after data analysis in any of our samples. Flongle sequencing had not been used before on such low contamination loads in food. However, it had been described for the detection of food viruses in plants for routine use (Liefting et al., 2021), but without indication of the contamination load. We believe that the contamination load in our samples is too low for Flongle technology to obtain sufficient reads. It has been acknowledged by ONT as an instrument that does not perform as efficiently yet as the MinION technology (already available for a longer time) and has pores that are more sensitive towards potential artefacts (e.g. the use of glass vials instead of plastic vials is required to not impact the sequencing). Therefore, a full characterization of the genome of the virus in contaminated food samples with Flongle sequencing is too challenging. However, Flongles might be optimized by the manufacturer in the future and could then be used for more complex cases.

As a final test for our initially envisaged open approach, we wanted to analyze food samples contaminated with another virus e.g. hepatitis A virus. We worked with a previously published dataset of celery spiked with norovirus and hepatitis A virus (Yang et al., 2017). This dataset was produced with an improved RNA extraction aiming at increasing the viral load in the extract by using ultracentrifugation and a commercial viral RNA extraction kit. Moreover, it was sequenced on an Illumina MiSeq. At the time of the publication, the authors were able to detect the viruses present in all the samples, even when two strains of norovirus were spiked in the same sample. After analysis with our bioinformatics work-flow (revised for Illumina reads), we could detect norovirus in 11 out of 13 samples spiked with norovirus, and HAV in 4 out of 4 samples spiked with it, with a completely open approach. We then built a consensus sequence for these strains, that could further be typed, but also placed in a phylogenetic tree. This goes beyond the analysis previously conducted on this dataset, and also the results that can be obtained with the currently available conventional methods. Notably, at the time of the publication of these sequences, a very open profiling method, Kraken, did not detect the viruses, as reported by the authors. Five years later, we could detect these viruses with the same tool, probably due to the update in the databases within this time period and the assembly of the reads prior to the taxonomic classification. Our analysis was able to detect two different viral species when co-spiked in the same celery sample, which could then be characterized to the genotype level. Because our analysis workflow was based on the best hit with Mash, we could however not separate two strains of the same genogroup, and the database used (Refseq) only contained on reference for norovirus GII. Previously, we had shown that a reference-based mapping tool such as Metamaps (Dilthey et al., 2019) for long reads allowed to separate closely related bacterial strains in the same food sample (Buytaers et al., 2021b). However this tool did not give satisfactory result in this case study (data not shown) because of the shorter read length, low contamination dose and low abundance of viral sequences in the associated database. A follow-up bioinformatics study might be able to find more specific tools to attain strain level for closely related strains of the same genogroup, at very low contamination level, in the same sample. However, the focus of this paper was to deliver a proof of concept at the wet-lab level and to obtain relatedness using a phylogenetic analysis, which is not possible with the conventional methods. Nevertheless, the analysis of the public dataset from Yang and colleagues allowed us to prove that our analysis workflow was performant for samples spiked with levels as low as 10^3 genome copies. This improvement in detection level is probably due to the targeting of viral particles prior to the RNA extraction step. The sequencing technology presumably had no impact on the results as fewer and shorter reads were produced with Illumina sequencing.

In conclusion, this study aimed at investigating which approach would be appropriate for further formal validation to be used for foodborne viral detection and full-genome based characterization. However, some further development would still be necessary before applying it in routine laboratories, as our results were not all positive and highlighted the complexity of such experiments when a virus is present at low dose in a sample. Some lessons learnt from our experiments with low contamination in food samples were that several methods that had been reported to give good results on higher contamination loads (e.g. clinical samples) did not work with our samples (i.e. poly(A) capture, rRNA depletion). Nevertheless, other methods we applied, were able to characterize the norovirus spiked in food samples: notably, the shotgun metagenomics methods on cDNA (method C) or amplified cDNA (method D) allowed to obtain a consensus sequence covering 85 to 99% of the genome in the samples spiked at the highest concentration (10^7 genome copies in 25 g of fruits). For medium contamination dose (10^5 genome copies in 25 g of fruits), a targeting approach such as SureSelect (method F) gave even better results, although it is very time-consuming, costly and doesn’t allow for an open approach. Therefore, these methods that gave positive results in our study still have limitations. In the future, if improved, the adaptive sampling proposed by ONT could be a cheaper alternative that could also target more than one pathogen. For lower contamination doses, our developed bioinformatics workflow was able to detect and characterize norovirus and hepatitis A virus at doses as low as 10^3 genome copies in 50 g of matrix, but with RNA extracted with another method after ultracentrifugation to enrich viral particles prior to extraction (Yang et al., 2017). Although we initially thought that using the currently accredited RNA extraction protocol would be the easiest way to later implement this new approach in routine, as an alternative with access to an ultracentrifuge potentially not possible for all reference laboratories, and a pre-enrichment step being more time-consuming, we show in this study that there is a trade-off between straightforward applicability and the potential limit of detection. Notably, this limit of detection, i.e. the sensitivity, and the specificity and reproducibility of the method still have to be determined in follow-up validation studies while this work only investigated the possibility to obtain whole genome characterization and phylogeny at a few different contamination loads as a proof of concept. These validations are not as common for metagenomics (Peterson et al., 2022) as they are for qPCR tests, given the cost per sample. So far, there is no consensus on how to conduct these validations or how many samples are necessary (Negida et al., 2019), and in the case of norovirus, access to references materials spiked at various contamination levels will prove challenging as we are bound to the genome copies present in the spiked lenticules. Moreover, the costs and efforts of adapting the ISO-based routine sample preparation in reference laboratories should be carefully evaluated against the benefits obtained when using an improved RNA extraction protocol. Another limitation to this method, because it is based on nucleic acids, is the possible characterization of a pathogen in a non-infective state (not living). However, we believe that when a person was infected by ingesting contaminated food, it is important to find the source of the disease even if the pathogen is not infective anymore. Therefore, our focus was to obtain relatedness between cases using phylogenetic trees, for which a nucleic acid-based method proved successful. In addition, most of the metagenomics workflows including the one presented in this study still rely on the use of command line and scripts, which is not straightforward for non-experts and prevents its use in routine, and should be addressed in the future. The databases that are used for the bioinformatics analysis should also be continuously updated and completed, especially for the improvement of investigations of mixed datasets such as the metagenomics ones. It is however important to note that in all the cases where a characterization of the virus was possible, we were able to obtain a genome which could then be compared to other cases by phylogenetics, which goes well beyond the results obtained with the current methods of analysis of norovirus in food. This paper aimed at sharing some lessons learnt, including approaches that failed for our samples. We believe that this contribution is also important for the scientific community to grasp information on what should not be repeated in the future, and from where to build further in a community effort. Above all, metagenomics is still a new approach and necessitates proofs of concepts such as this one to advance the field, as was requested by EFSA (EFSA, 2019b).

Acknowledgements

We appreciate the help of Assia Saltykova and Alexander van Uffelen for their advices on data-bases and bioinformatics tools. We are grateful to the service of food pathogens (Sciensano) and their technicians for the sharing of their RNA laboratory, their training for the specific protocols of norovirus RNA extraction and detection, and for the qPCR performed on the spiked samples. Finally, we acknowledge Stefan Hoffman for setting up ONT sequencing in the laboratory of Transversal Activities in Applied Genomics (Sciensano).

Supplementary data:

Supplementary Material S1 can be found online at https://www.mdpi.com/2304-8158/11/21/3348/s1?version=1666687864

Data availability statement:

all data is publicly available under BioProject PRJNA878666.

CHAPTER 8
General discussion, conclusion and perspectives

The current practices for the detection and characterization of microbiological food contaminants in the respective Reference Laboratories are based on a series of tests. These depend of the contaminant, but generally a screening is performed on the food, often via qPCR. If a marker indicates the presence of a contaminant, this is investigated further and the gold standard, if possible depending on the contaminant, would be to obtain an isolate and characterize it. This is effective, but an isolate is not always obtained, and it is time-consuming as it requires growth for about a day on each media. Therefore, some specific case studies might slip through the cracks of this testing. Indeed, in 2020, 39,8% of foodborne outbreaks in the EU were caused by an unknown agent, as it could not be characterized in the analysed leftover food. Moreover, it has been shown in some foodborne outbreaks that more than one pathogenic strain was present (Kinnula et al., 2018; Somerville et al., 2018), but these cases of co-infections might be underestimated due to the isolation process, and hence remain undetected. Moreover, in the case of GMMs, which might be difficult to grow due to auxotrophy, and when the modified genome has not been previously unravelled, it is tedious to fully characterize the contaminant (Fraiture et al., 2020e) or sometimes even unattainable. In the case of viral pathogens, culture enrichment is very arduous (Jones et al., 2015) or even impossible, and the characterization with the available methods is incomplete (Bosch et al., 2020; Desdouits et al., 2020a). A new method allowing detection and characterization of biological contaminants without *a priori* isolation, i.e. metagenomics, had started to show some promising results when this PhD research started. Therefore, the objective of this PhD was to investigate how metagenomics could potentially resolve the issues encountered with the conventionally used detection/characterization methods for foodborne microbiological contaminants, while obtaining at least the same level of information. This means the detection of the microbiological contaminant/pathogen, its identification and characterization (i.e. dependent on the contaminant, serotype/serovar/genotype, detection of AMR and/or virulence genes, unnatural associations) and the ability to relate several cases in a phylogenetic tree. Finally, we also aimed at developing a workflow both at the wet laboratory and the dry laboratory (bioinformatics) levels, potentially applicable in a routine setting at national reference laboratories. This was done using several case studies representing various matrices with various contaminants at different levels i.e. a bacterial pathogen at a low contamination level in a complex food matrix, after enrichment (Chapters 3-4-5), a bacterial GMM in a non-complex matrix without enrichment (Chapter 6) and a viral pathogen at a low contamination level in a complex food matrix without enrichment (Chapter 7). These case studies were investigated with short and/or long reads sequencing, which also influenced the data analysis tools to use.

The detailed results of this PhD research were discussed separately and extensively in each chapter (Chapters 3 to 7). Therefore, the present chapter (Chapter 8) places these findings in a broader view to answer the scientific questions raised in this PhD research (Chapter 2).

8.1. Which metagenomics approach could allow characterization and relatedness at least at the same level as the conventional methods?

Metagenomics allows to study all organisms present in a sample at once and therefore to obtain information from the contaminant in the food being analysed without the need for isolation (Escobar-Zepeda et al., 2015; Forbes et al., 2017). In the scientific community, two different methods have been presented to obtain information without isolation of the contaminant. They deliver different levels of information but are based on the DNA or RNA extracted from the whole sample: i.e. metabarcoding (Woese and Fox, 1977), based on the sequencing only of the 16S rRNA or other targeted regions, in order to differentiate the species present in the sample, and shotgun metagenomics (Stein et al., 1996; Handelsman et al., 1998), based on the sequencing of all the DNA or RNA. These were presented in the introduction (Chapter 1) of this thesis. For our research, we have decided to use shotgun metagenomics, as it was the only approach allowing us to obtain the entire genetic information from the sample and therefore potentially obtain the entire genome of each strain in the sample. This genome could then be characterized at the same level as the conventional methods.

Our workflow aimed first at detecting the possible contaminant in the sample without *a priori* knowledge, by looking at all the sequenced DNA (long or short reads) originating from metagenomics sequencing of the total sample nucleic acid extraction (using extraction methods as close as possible to what is used in routine). Then, the first step of the data analysis is to perform a profiling or taxonomic classification step. Thereafter, our goal was to distinguish each strain in the sample so that it could later be characterized at least at the same level as would be possible when using isolate WGS data, e.g. by detecting different genes present in the inferred genome/strain. Finally, we also intended to conduct a study of relations between cases based on a phylogenetic analysis of the obtained strains. Based on this, we would obtain at least the same level of information as the conventional methods. This workflow was evaluated using different case studies, representative of possible scenario in routine settings. If needed, specific adaptations were made to the workflow, depending of some specific characteristics represented by the case study (cfr. following sections, Figure 8.1).

The profiling step performed in this work was mainly conducted using the classifier Kraken2 (Wood and Salzberg, 2014). It was accessible on the Sciensano in-house galaxy instance and server, with in-house pulled databases based on Refseq. The first studies (Chapters 3-4-5-6, studies of bacterial pathogens and GMMs) were performed using two databases in a stepwise fashion: first we investigated with a mammal database, to remove the host (matrix) reads, then a database of archaea, bacteria, fungi, human, protozoa, and viruses was used to characterize the reads that were not classified in the first step. In a next phase, (Chapter 7, viral RNA pathogens), the two databases were merged to have only one database called “full”. Using one database helps to avoid misclassification that could have occurred in

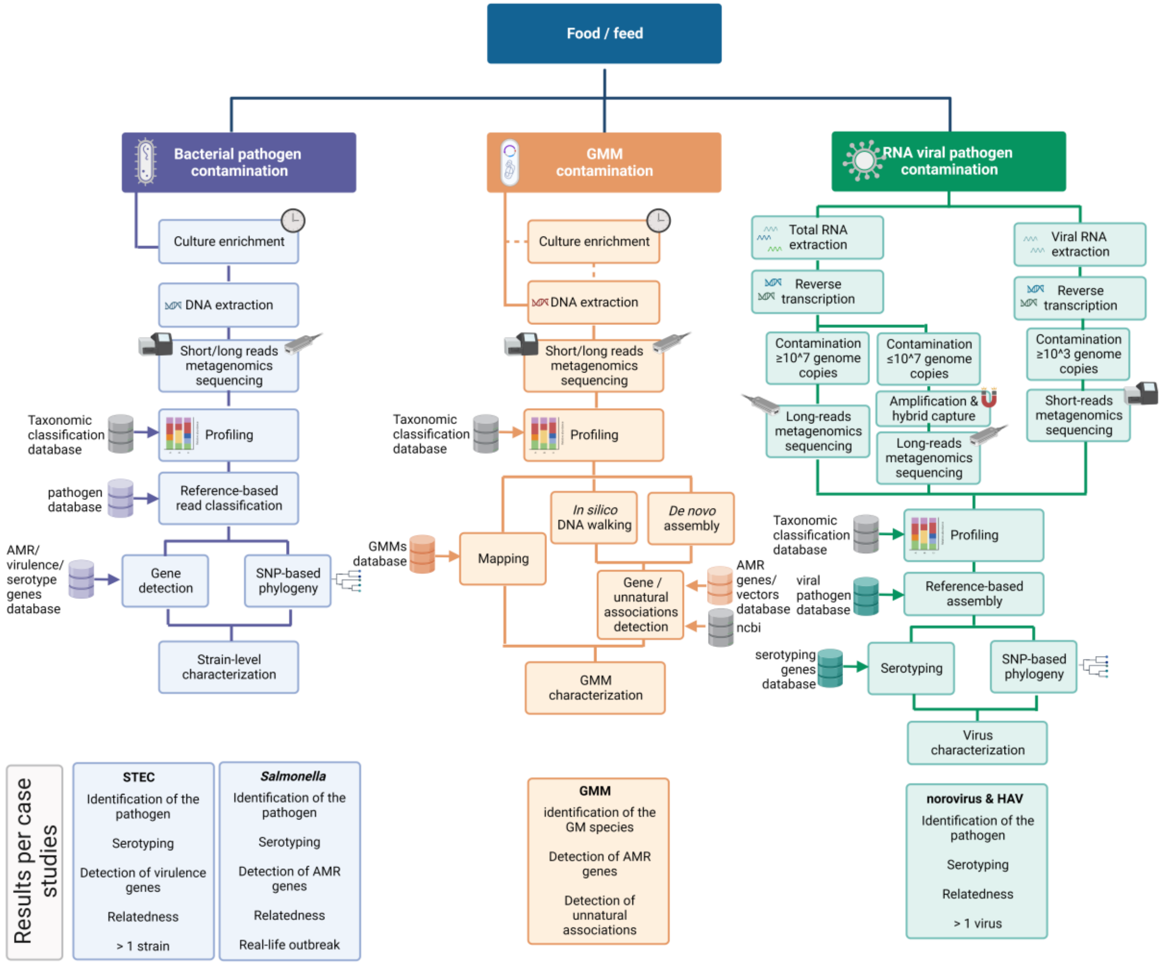

Food / feed
Bacterial pathogen contamination
Culture enrichment
DNA extraction
Short/long reads metagenomics sequencing
Taxonomic classification database
Profiling
pathogen database
Reference-based read classification
AMR/ virulence/ serotype genes database
Gene detection
SNP-based phylogeny
Strain-level characterization
GMM contamination
Culture enrichment
DNA extraction
Short/long reads metagenomics sequencing
Taxonomic classification database
Profiling
GMMs database
Mapping
In silico DNA walking
De novo assembly
Gene / unnatural associations detection
AMR genes/ vectors database
ncbi
GMM characterization
RNA viral pathogen contamination
Total RNA extraction
Reverse transcription
Contamination ≥10^7 genome copies
Contamination ≤10^7 genome copies
Long-reads metagenomics sequencing
Amplification & hybrid capture
Long-reads metagenomics sequencing
Viral RNA extraction
Reverse transcription
Contamination ≥10^3 genome copies
Short-reads metagenomics sequencing
Taxonomic classification database
Profiling
viral pathogen database
Reference-based assembly
serotyping genes database
Serotyping
SNP-based phylogeny
Virus characterization
Results per case studies
STEC
Identification of the pathogen
Serotyping
Detection of virulence genes
Relatedness
> 1 strain
Salmonella
Identification of the pathogen
Serotyping
Detection of AMR genes
Relatedness
Real-life outbreak
GMM
identification of the GM species
Detection of AMR genes
Detection of unnatural associations
norovirus & HAV
Identification of the pathogen
Serotyping
Relatedness
> 1 virus

Figure 8.1: General recommendations of workflow to follow (wet-lab and dry-lab) for each case study, based on the results presented in this thesis. *Dotted lines: optional culture enrichment for GMM contaminated samples. Workflows for pathogen contaminations in the context of outbreaks. Figure made with Biorender.com*

the previous two-steps taxonomic classification (false positive classified as mammals cannot be classified as bacteria). This required increasing the computing power necessary for the analysis with such a big database. It is important to note that although classifiers can be compared and benchmarked, which is a study on its own, no one-size-fits-all approach exists, and different tools and databases will give different results (Chapter 4, Salmonella outbreak study and Chapter 6, GMM study, Wright et al., 2022).

The strain-level (i.e. aiming for the same level of analysis as with the conventional methods involving isolate WGS) bioinformatics analysis used in this work was developed during the PhD of Assia Saltykova (Saltykova, 2022). After comparison of the possible tools to bioinformatically separate the genomes of the strains present in the shotgun metagenomics dataset, it was shown that a reference-based read classification was the most appropriate approach for the sequencing depth and the contamination levels expected in food samples (Figure 8.1). To this end, Saltykova et al. compared the results obtained with two tools, Sigma (Ahn et al., 2015) and Sparse (Zhou et al., 2018), on a sample containing non-pathogenic *E.coli* spiked in vivo with STEC, and several *in silico* spiked food samples (Saltykova et al., 2020). Sigma gave better results for the detection of virulence genes in the clustered reads corresponding to the STEC strains, and this tool was used in this PhD for the analysis of short reads data of STEC spiked food samples (Chapter 3). As we designed an open approach, we showed that it could be applied to another pathogen after enrichment if the database was adapted after determination of the pathogen during the profiling step. This was done on the case study of *Salmonella* outbreak in Chapter 4. A similar analysis was conducted on long reads sequences of the STEC spiked samples (Chapter 5), using Metamaps (Dilthey et al., 2019) for the long reads assignment instead of Sigma (Figure 8.1). When the analysis was transposed to the study of GMMs (Chapter 6) and viral RNA pathogens (Chapter 7), the bioinformatics workflow had to be adapted (Figure 8.1). This will be elaborated further in section 8.3.

8.2. How does the contamination level and/or the matrix influence the approach to be followed?

In the work presented, the selected case studies represented various food matrices (minced beef meat, cheese, a meal composed of potatoes, tartar sauce and fish, vitamin powder, raspberries, shellfish), with various contamination levels i.e. low level of contamination followed by culture-based enrichment for STEC and Salmonella (Chapters 3-4-5), medium and low level of contamination with no culture-based enrichment in a simple or complex matrix for GMM (Chapter 6) and norovirus (Chapter 7) respectively. These matrices and contamination levels were representative of the samples that could be received at the NRLs, as exposed in each chapter.

The nature of the matrix only had a small impact on some specific methods to use to prepare the genetic material (e.g. fat removal for the cheese, pH adjustment for the raspberries). Essentially the same sample preparation method could be followed for the different matrices except for these small variations: sampling of a fraction of the contaminated food matrix, homogenisation (also to homogenize the contamination that can be heterogeneous) and extraction of the genetic material. Therefore, as shown in the different chapters of this thesis, the metagenomics method could be successfully transposed to other matrices in order to characterize contaminants in any type of food product.

The contamination level, however, had a strong influence on the sample preparation method and analysis workflow to be followed. For the sample preparation, we demonstrated that after bacterial culture enrichment using appropriate medium and temperature, the low contamination of STEC (5 CFU/25g) and *Salmonella* (Chapters 3-4-5) was easily detected and characterized after DNA extraction with a commercial kit. This was possible even when two strains of the same species were present in the sample. DNA amplification was tried in order to improve the results, but we showed that this was not necessary when culture enrichment could be performed on the food sample. In the case of a non-complex matrix and no enrichment (GMM contamination, Chapter 6), most of the extracted DNA corresponded to the contaminant. Therefore, except for the DNA extraction, no sample preparation was necessary. The extracted was conducted with a commercial kit (Nucleospin food, the same kit as used with culture enrichment of bacterial pathogens i.e. Chapters 3-4-5). However, other DNA extraction methods could be tested in further studies to try and increase the read length for long reads sequencing and therefore allow detection of the unnatural associations directly in the reads without assembly, or to obtain a strain-level characterization if several GMM strains are present in the sample (Figure 8.2), a case that was not tested in our work (Figure 8.1). However, obtaining long DNA fragments might not be possible due to the pre-treatment of the microbial fermentation products (e.g. vitamin powder) before it is sold on the market, aiming at removing any genetic material or contamination from the product. Finally, when the culture-based enrichment was not possible, and the contamination was low in a complex matrix, such as the case of the norovirus spiked samples (Chapter 7), we showed that the virus could be detected and characterized with a classical shotgun metagenomics, when 10^7 genome copies were spiked (Figure 8.1). It was previously shown that very few reads could be assigned to norovirus after shotgun sequencing of other food samples (Bartsch et al., 2018). The whole transcriptome amplification did not improve this result. However, we could obtain the viral genome in samples spiked with 10^5 genome copies when a target enrichment was used (hybrid capture using SureSelect). This method had to be preceded by the whole transcriptome amplification to have sufficient cDNA input for the protocol. We also showed that the use of RNA extraction methods targeting the virus might circumvent the need for this targeting (Conceição-Neto et al., 2015; Yang et al., 2017) and allow to obtain a characterization of the viral pathogen from 10^3 genome copies (Figure 8.1).

Concerning the impact of the contamination level on the bioinformatics approach, we decided to rely on reference-based classification methods in the studies performed with food pathogens (Chapters 3,4,5,7) as described by Saltykova et al. (Saltykova et al., 2020), both for the long and the short reads. This workflow also allowed us to characterize two strains spiked at very low level (5 CFU/25g) in the same food sample. However, the characterization of one of the strain was incomplete. This could be resolved by having more genomes in the database used to classify the reads to each strain. The above-mentioned approach could not be used to study artificial constructs such as GMMs due to their unnatural nature, which will be explained in more detail in section 8.3.1. The data analysis also had to be adapted when studying viruses, which will be explained in section 8.3.2. This was mainly due to the underrepresentation of viral pathogens in some databases, but also due to the very low level of contamination which required to decipher between real detection of a contaminant or false positive (or negative) during the profiling step. For this reason, a *de novo* assembly was conducted before the taxonomic classification. Nonetheless, we showed that we could characterize one or two different viruses present in the same sample at a level of contamination as low as 10^3 genome copies per 50g of food matrix (Figure 8.1).

Based on the case studies investigated in this work, we could conclude that when a culture enrichment is possible, shotgun metagenomics seems able to attain strain-level with reference-based analysis methods and without specific sample preparation (Figure 8.1) on different matrices contaminated at very low level (a few CFUs). However, when this culture enrichment is not desirable nor possible (e.g. case of GMMs, chapter 6, and viruses, chapter 7), shotgun metagenomics can be used as such only if the contamination level is high enough. When the contamination level is low, a targeting such as presented with SureSelect for norovirus is necessary (Figure 8.1), but this is a targeted method so it will only detect the virus of interest and it has not been commercially developed for all possible contaminants. The adaptive sampling proposed during ONT sequencing could replace this targeting if we would obtain sufficiently long DNA fragments for such method to be effective. Other DNA/RNA extraction methods should be envisaged to obtain high molecular weight genetic material, and adaptive sampling still has to be proven effective to attain strain-level for this case study, with the adapted database (Figure 8.2). Alternatively, the amplification of the genetic material did not improve the results we obtained, but can be necessary to have sufficient genetic material input for some protocols, such as the SureSelect (hybrid capture). It might also be used before ONT sequencing when the DNA/RNA concentration is very low, as this sequencing technology requires high amounts of genetic material (~1µg depending on the flow cell). However the amplification might also introduce bias (some strains could be preferably amplified compared to others) as well as branching in the amplified DNA, which will block the sequencing pores of the ONT flow cell. Therefore, a de-branching with the T7 enzyme is proposed, but could result in fragmented DNA (Oxford Nanopore technologies, 2019). In future studies, this protocol could be tested and improved (Figure 8.2).

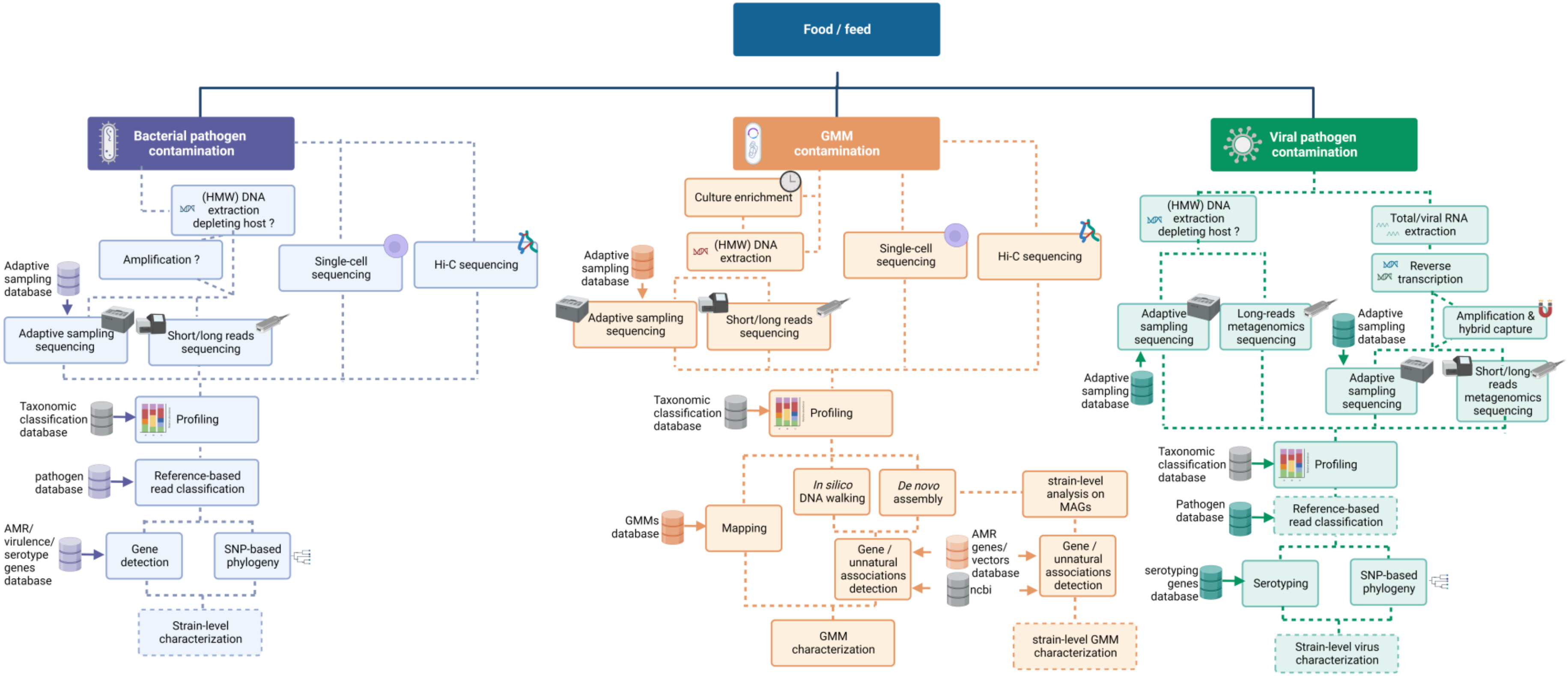

Food / feed
Bacterial pathogen contamination
GMM contamination
Viral pathogen contamination
(HMW) DNA extraction depleting host ?
Amplification ?
Single-cell sequencing
Hi-C sequencing
Adaptive sampling database
Adaptive sampling sequencing
Short/long reads sequencing
Taxonomic classification database
Profiling
pathogen database
Reference-based read classification
AMR/ virulence/ serotype genes database
Gene detection
SNP-based phylogeny
Strain-level characterization
Culture enrichment
(HMW) DNA extraction
Single-cell sequencing
Hi-C sequencing
Adaptive sampling database
Adaptive sampling sequencing
Short/long reads sequencing
Taxonomic classification database
Profiling
In silico DNA walking
De novo assembly
strain-level analysis on MAGs
GMMs database
Mapping
Gene / unnatural associations detection
AMR genes/ vectors database
ncbi
Gene / unnatural associations detection
GMM characterization
strain-level GMM characterization
(HMW) DNA extraction depleting host ?
Total/viral RNA extraction
Reverse transcription
Adaptive sampling sequencing
Long-reads metagenomics sequencing
Adaptive sampling database
Amplification & hybrid capture
Adaptive sampling database
Adaptive sampling sequencing
Short/long reads metagenomics sequencing
Taxonomic classification database
Profiling
Pathogen database
Reference-based read classification
serotyping genes database
Serotyping
SNP-based phylogeny
Strain-level virus characterization

Figure 8.2: Perspective of methods that were discussed but not yet tested during this thesis nor in other studies and could potentially improve the results and/or the time to results of a metagenomics approach for various types of microbiological contaminants in food. *Dotted lines: not yet tested methods, not yet proven resolution. Figure made with Biorender.com*

8.3. How to adapt the approach depending on the microbiological contaminant?

8.3.1. How to adapt the approach for the analysis of GMMs?

We investigated the use of shotgun metagenomics for food matrices contaminated with pathogenic bacteria, involving a culture-based enrichment (chapter 3-4-5). The case of the contamination of vitamin powders with a GMM (Chapter 6) however, was representative of a low complexity matrix which implied that most genetic material extracted from the sample corresponded to the contaminant. No enrichment was conducted, but this culture might also not be feasible as GMMs are not always easy to culture due to the possible presence of genetic modifications requiring specific growth factors for the microorganism to multiply (auxotrophy). These modifications are added as a safety net to hinder the proliferation of the GMM outside the production conditions. The DNA was extracted with the kit also used for the study of bacterial pathogens (i.e. Nucleospin food), and used at the NRL.

For the bioinformatics analysis, the detection and characterization of a GMM in a food/feed sample was slightly different to the analysis of pathogens. Indeed, after the profiling step, the focus is not the investigation of some markers of virulence or type, but the detection at strain level of ARGs and the discovery of one or several unnatural associations or junctions within a wild type genome. Therefore, the bioinformatics analysis workflow had to be adapted (figure 8.1), i.e. after determining the species or genera present in the sample with taxonomic classification, we looked at the possible presence of AMR genes or a known shuttle vector, reported in several GM constructs. This was done after *de novo* assembly of the reads as no reference could be used per se because we were investigating unknown artificial constructs. Finally, the contigs harbouring these markers were associated to the closest wild-type genome in the nucleotide database using blast. We could then determine if a contig was harbouring a mix of genomes of several species, a sign of an artificial construct. Finally, as a validation of our method and because we used *de novo* assembly, which could create chimeras and therefore lead to the detection of false unnatural associations, we confirmed our findings by mapping the reads to a known GMM reference genome possessing the same genes and shuttle vector. However, such reference genomes of GMMs are not made publicly available by the producing company, preventing to conduct this analysis when the GMM has not been previously characterized. Obtaining longer DNA fragments for long reads sequencing would help for future analyses, in particular of unknown GM constructs. Moreover, all known/already characterized GM constructs should be centralized in a publicly available database to facilitate the characterization of GMM strains using metagenomics. Finally, our study analysed samples containing only one strain of the same species, but the workflow should be adapted in order to allow the analysis of several strains, possibly several genetically modified microorganisms (Figure 8.2). For this purpose, several bioinformatics analyses could be tested, but a new study showed that investigating the depth of coverage of certain regions in the metagenome-assembled genomes (MAGs) might suggest the presence of different strains in the sample (D'aes et al., 2022), however this still requires substantial manual analysis.

8.3.2. How to adapt the approach in case of bacterial or viral pathogenic contaminant?

Besides case studies involving bacterial contaminants (chapters 3-4-5-6) we also included a case study of viral RNA food contaminations (norovirus and hepatitis A virus) (chapter 7).

At the sample preparation level, most foodborne viruses, including norovirus and HAV, cannot be easily cultured, and therefore no enrichment was possible for our work with viruses. We already worked without enrichment for the detection and characterization of GMMs in a microbial fermentation product (chapter 6). But microbial fermentation products are rather 'simple' matrices in terms of generating background reads. For viruses in complex food matrices, however, the contaminant's load within all extracted nucleic acids of the sample is limited (as shown in Chapter 7). Therefore, in our case study, six sample preparations methods were tested, all after following the standardized (i.e. used in routine in the NRL) total RNA extraction protocol on the different samples (ISO 15216-2). As presented in the previous section, the protocol to be used will depend on the viral contamination level (Figure 8.1). Moreover, depending on the choice to modify or not the RNA extraction from the ISO (so the ease of transposing the method in the routine setting), a lower viral contamination level might be characterized with shotgun metagenomics. Importantly, we decided to work with DNA sequencing technologies and had to reverse transcribe the extracted RNA in all cases before sequencing. Some direct RNA sequencing methods have been previously used with ONT as well (Wongsurawat et al., 2019), but these produce a much lower data output. As the detection and characterization of the virus was already challenging with the output of a DNA sequencing, we did not try this other approach. However, future improvements in this technology might allow for this approach to become applicable for RNA virus detection using shotgun metagenomics. If studying DNA viruses (Figure 8.2), such reverse transcription would not be necessary. Moreover, the use of a host depletion kit such as the HostZERO (removing eukaryotic DNA) could be an interesting alternative to the total DNA extraction that includes a lot of genetic material from the matrix. However, this kit would have to be tested for this application. Moreover, if the HostZERO extraction method could produce sufficiently long reads, adaptive sampling might be used as an alternative to the targeting of the virus with SureSelect (and a more open approach for several different viruses depending on the database used while SureSelect targets only one virus).

At the level of the bioinformatics analysis, we showed in the study of the *Salmonella* outbreak (Chapter 4) that for the study of a different bacterial contaminant, the same workflow could be used except for the modification of the database to use for the strain-level inference of the reads. For the study of viruses, however, this workflow had to be modified. A reference-based mapping such as MetaMaps (for long reads sequencing) could not be successfully used to attain strain-level characterization of the contaminants in our samples

(Chapter 7) because of the very low amount of reads that were classified as pathogenic virus after using the tool, due to the low representation of virus references in the database associated with the tool. Therefore, another approach had to be developed for this case study, which allowed to attain strain-level characterization, even for two strains of different pathogenic viruses in the sample, but not for two strains of the same virus. A strain-level characterization of several strains of the same species was not possible for this case study because of the design of the bioinformatics workflow (as a reference-based assembly was conducted and the tool to determine the best reference to use – Mash - would only consider one best reference per species). It might be possible to develop a workflow attaining strain-level (Figure 8.2) but this might be challenging due to the very low contamination load in the samples, and it would require a database containing a large variety of virus reference genomes.

Notably, if studying DNA viruses, the same challenges would be faced as the difficulty resides in the low contamination load (without enrichment) and the poor representation of viral references that impede the use of MetaMaps for read classification. The bioinformatics workflow presented for RNA viruses could, however, still be used to study DNA viruses as such (Figure 8.2).

8.4. How does the sequencing technology influence the results?

Two sequencing technologies have been tested in this work: the short reads (i.e. 2x 250 bp) Illumina sequencing on an in-house MiSeq instrument (up to 15 Gb output), and the long-reads (> 1000 bp) ONT sequencing on Flongles (up to 2.8 Gb output), MinIONs (up to 50 Gb output) and GridION (5 MinION sequencing in parallel, allows adaptive sampling). The data from our work on the *Salmonella* outbreak (Chapter 4) was only generated with Illumina, we compared the use of the two technologies on different case studies (bacterial pathogen in Chapters 3 and 5 and GMM in Chapter 6), while we only used the nanopore technology for the study of viruses in food samples (Chapter 7). These technologies have been tested and compared as they produce short versus long reads with different degrees of error rate (0,01% error rate for Illumina, ~6% error rate for ONT) and hence will impact the tools and parameters to use for the data analysis (table 8.1). These technologies can produce sequencing data in real-time (ONT) or over a longer time frame (~48h, Illumina), and they come at a different costs and sequencing outputs depending on the number of samples to be analysed (Table 8.1). Moreover, both technologies have other requirements in terms of sample preparation: Illumina only requires 1 ng of fragmented DNA while ONT requires 1 µg (or more, depending on the reagents) of high molecular weight DNA (shorter DNA fragments can be sequenced but will produce short reads).

In the case of the bacterial pathogens after enrichment, we showed that we could obtain a similar level of characterization at strain level after sequencing culture enriched samples spiked with STEC with Illumina or ONT (Chapter 3 and 5). With ONT sequencing, we could achieve the same characterization by developing a similar data analysis workflow specific for long reads, by replacing Sigma (Ahn et al., 2015) used for short read data by Metamaps

(Dilthey et al., 2019) for the strain-level reads assignment. In that regard, we concluded that the sequencing technology did not impact the results for a characterization of the pathogen at the strain level. However, we also showed that we could conduct another type of analysis on long reads without the bias of a *de novo* assembly: the *in silico* DNA walking. This approach consists in the determination of the genomic context surrounding a gene of interest on long reads. This allowed us to associate the *stx* genes with the *Escherichia* genome, confirming rapidly that a STEC strain was present in the sample, and could be used for a screening using the low-cost Flongle flow cells.

In the case of the GMM contaminations (Chapter 6), we also noted that we could obtain the same level of characterization with the two sequencing technologies. The *de novo* assembly conducted with SPAdes (Bankevich et al., 2012) on short reads was then replaced by Canu (Koren et al., 2017) for the reads obtained with ONT sequencing. One promise from the long reads sequencing would be to conduct *in silico* DNA walking directly on long reads and detect unnatural associations or determine if the full length of the AMR gene is present in the DNA. This would allow to skip the *de novo* assembly step that can create chimeric contigs. However, the reads produced from a processed matrix such as the microbial fermentation products, without enrichment were not very long. It might be improved in the future by testing other DNA extraction methods to use this technology at its full potential (Figure 8.2). However, it might be impossible to obtain high molecular weight DNA from this kind of matrix. Moreover, in contrast to the use with an enriched matrix, the use of the Flongle flow cell was not adapted to the complexity of this GMM case study as no *de novo* assembly could be conducted due to the very low coverage of the data.

Finally, ONT was used for the analysis of viral contaminations in food because of cost-effective reasons (cfr. section 8.5). We showed that we could obtain an almost complete genome and characterize it when sequencing a sample with sufficient contamination load on a MinION flow cell. The Flongle flow cell did not perform well for this case study of low contamination in a complex matrix. We didn’t use short read sequencing for our spiked samples in this case study. However, we used our data analysis workflow to analyse a previously published dataset (even lower viral contamination in food, for which another RNA extraction procedure than ours was used) generated on Illumina MiSeq (Figure 8.1). We could obtain an equal amount of characterization level as we did for the ONT-based workflow. Therefore, we believe that the difference we see in the level of contamination at which we could characterize the virus is not linked to the sequencing technology, but to the RNA extraction instead.

In conclusion, these studies showed that both technologies provide similar levels of information on the contaminant present in the food. The choice between the technologies depends on the amount of samples to be sequenced in one run, as will be presented in the next section. ONT is now overcoming the high error rate that it used to be assimilated to with the latest pores, reagents and basecalling tools, and it has the advantage to allow a rapid analysis directly on the long reads, circumventing the need for an assembly (*in silico* DNA walking, Chapters 5-6). The Flongle produces a smaller output that can be used for a screening of the samples. However, ONT is a newer technology and therefore still requires more follow-up of new tools that are constantly being developed. Finally, Illumina recently announced the launch of long reads to be able to be generated on their ‘short read’ instruments (Illumina, 2022). It would be interesting to test this technology in the future for its application in shotgun metagenomics.

8.5. How to achieve fast and cost-effective results?

In the scope of the resolution of a *Salmonella* outbreak to its food source (Chapter 4), we could show that shotgun metagenomics is *per se* a method giving results faster than the conventional methods for bacterial contaminations, because it avoids the fastidious isolation step. Obtaining a bacterial isolate takes in average a week because it involves the succession of cultures on selective and unselective media, and a colony can only be clearly observed on a plate after 24 hours of enrichment (it can be more depending on the growth rate of the species). For our case study involving the *Salmonella* outbreak (chapter 4), we received the contaminated food samples linked to the outbreak afters they had been enriched for 18 hours (as stipulated in the ISO) and tested with qPCR for the presence of *Salmonella* (ISO: International Organization for standardization, 2017). We performed a DNA extraction of the entire samples, and could start an Illumina sequencing run the same week as the samples arrived at the NRL. After short read sequencing (2 days) and bioinformatics data analysis (1 day), the entire process from reception of the samples to the delivery of the results took one week. Moreover, we demonstrated that we were able within this timing to obtain the same information as the conventional method, therefore replacing them by only one test (Table 8.1): detection of the contaminant (*Salmonella* genus detected by taxonomic classification), recovery of the strain genome and characterization (to the serovar level, as well as detection of the AMR gene *aac(6’)- Iaa_1*) as well as relatedness to isolates from food and human origin of the same outbreak with 0 SNP difference, and clear separation to sporadic cases and another Salmonella outbreak. With the case study of the STEC spiked samples (Chapter 3), we demonstrated that the same workflow could be followed when more than one strain of the same species has to be characterized in the same sample.

Characteristic	Phenotypic tests			Molecular techniques			Metagenomics	
	Microscopic techniques	Immunological techniques	Other phenotypic tests*	based on PCR	not PCR-based such as PFGE	Whole genome sequencing	Illumina sequencing	ONT sequencing
Necessity for isolation	Not always +/-	Yes -	Yes -	Not always +/-	Yes -	Yes -	No ++	No ++
Time to result	5-30 minutes ++	10 min -1h (after isolation) +/-	Various (after isolation) +/-	1-3 hours ++	3 days (after isolation) -	3-4 days (after isolation) -	3-4 days +/-	1-2 days +
Sensitivity / resolution	1 cell/visual field ± 10^4 CFU/ml +/-	10^4-10^5 CFU/ml +/-	Various +/-	1 genome/PCR reaction +	+/-	SNP level ++	(depending on the contamination load, sample preparation and sequencing depth) +/-	(depending on the contamination load, sample preparation and sequencing depth) +/-
Specificity	-	+/-	-	+	+ (depending on the database)	++ (depending on the database)	+ (depending on the database and sequencing depth)	++ (long reads, depending on the database and sequencing depth)
Simultaneous multiparameter testing, including relatedness	+/-	-	- -	+	-	++	+++	+++
Differentiation of dead/viable cells (without taking isolation into account)	+/-	+/-	+/-	-	-	-	-	-
Reproducibility	+	+	+/-	+	+	+	+ (to be further validated)	+ (to be further validated)
Data analysis	+	+	+	+	+/-	Requires bioinformatics knowledge -	Requires expert bioinformatics knowledge --	Requires expert bioinformatics knowledge Few long-reads tools ---
Interpretation of results	Human bias possible -	+/-	Human bias possible, need for databases +/-	+	+/-	+/-	Not always strain-level -	Not always strain-level -
Labour intensity	+/-	+/-	- -	+	-	+	+	+
Cost of materials	-	+/-	+	+/-	-	-	-- (depending on number of samples)	- (depending on number of samples)
Investment of equipment	-	+/-	+/-	-	-	- -	- -	+/-

Table 8.1: Perceived characteristics of conventional and newer methods, including metagenomics as proposed in this thesis, for the characterization of microorganisms. **Other phenotypic tests described in the introduction (Chapter 1) e.g. API tests, maldi-TOF MS, GC FAME. The symbols describe 'overall perceived as a negative/positive/neutral intrinsic characteristic/(dis)advantage for the user', and more into detail: +: positive characteristic; + +: very positive characteristic; +++: extremely positive characteristic; +/-: average/neutral characteristic; -: negative characteristic; - -: very negative characteristic; ---: extremely negative characteristic. Figure adapted from Jasson et al. (Jasson et al., 2010)*

These studies were both performed using Illumina sequencing. However, this technology requires to sequence for 48 hours (when using the 2x250 bp sequencing on Illumina Miseq) and, because of the low number of samples that can be combined for a metagenomics run to obtain sufficient sequencing depth, this sequencing method is still expensive compared to the conventional methods and the WGS of isolates (Table 8.1). For a deeper sequencing, less samples would be placed on the cartridge and the price would be higher. These issues might be improved by using ONT sequencing. Indeed, ONT offers a relatively easy and fast protocol for the generation of the library, and a real-time sequencing for a very limited investment (the MinION instrument being about 100 times less expensive than for Illumina sequencing, while one MinION flow cell is half the price of one MiSeq cartridge, with, under perfect conditions, at the time of our studies, similar Gb outputs, Table 8.1). It allows to analyse 1 sample per flow cell, at a reasonable cost (Table 8.1). We compared the results that could be obtained with Illumina, ONT MinION or ONT Flongle sequencing on the same sample of minced meat spiked with STEC and enriched for 24 hours (chapter 5). We proved that a strain-level characterization could be obtained with both technologies, by using a similar data analysis workflow. For MinION sequencing, this was possible after only 12 hours of sequencing. After 24 hours of sequencing on a Flongle (1/10th of the price of a MinION flow cell for 1/10th of the amount of pores and therefore the expected sequencing output), it was also possible to obtain the same level of information if the DNA of the sample was extracted with the HostZERO kit, selectively depleting the eukaryotic DNA. Although Illumina and ONT offer similar results at strain level, the sequencing technology should be selected based on the amount of samples to sequence in one run. Indeed, Illumina sequencing is more cost-effective for several samples, and we presented in the chapters on STEC contamination, Salmonella outbreak and GMM investigation (Chapters 3-4 and 6) that we could obtain good results with 12 samples in one run. ONT MinION sequencing can be successfully used when only one sample has to be sequenced and in order to obtain results quickly. In the case of norovirus contamination, we decided to sequence only with ONT because the amount of samples to sequence in an Illumina run would not be obtained in practice, and this would therefore prevent transposing this method into routine. It is necessary to keep in mind that the cost of the sequencing technologies remains higher than the cost of any single conventional test (Table 8.1), but the sequencing gives in one test the information obtained from multiple other methods combined, and the price has drastically decreased and is expected to continue to decrease, which could make metagenomics more accessible.

Moreover, our work on nanopore sequencing allowed us to develop an *in silico* DNA walking method. This approach is a faster way to confirm if a contaminant is present in a sample compared to a detailed strain-level metagenomics analysis, by looking at the genomic context surrounding a gene of interest on long reads to verify the species hosting this gene. Although it does not allow a strain-level resolution or a relatedness analysis, it supports a more accurate discrimination in the contaminants detected in the sample (pathogen or not, genetically modified or not). We presented a confirmation of the presence of a STEC after only 1h of sequencing on a MinION flow cell (Chapter 5). The same information could be ontained after one day of sequencing on the low-cost Flongle flow cell. Therefore, Flongle could be used as a low-cost screening method with *in silico* DNA walking, with the possibility to obtain strain-level characterization and relatedness if the sequencing depth allows it and if using a host depletion DNA extraction, as presented in our work on STEC (Chapter 5).

8.6. How to implement this approach in routine analyses? Is the sensitivity of the method sufficient to comply with the current routine practices?

The focus of this work was to obtain an alternative approach for the detection and characterization of food contaminants using metagenomics, that could be implemented in routine in the future. Therefore, our sample preparation and DNA/RNA extraction methods always stayed as close as possible to the current protocols in the reference laboratories, and the international standards.

We showed in our work on enriched bacterial pathogens (Chapter 3-4-5) that shotgun metagenomics can be used to characterize bacterial foodborne contaminants, even when more than one pathogenic strain was present, and to resolve an outbreak to its food source without ambiguity. This was done after enrichment in a non-selective buffer, following the protocol already used in the routine laboratories (ISO: International Organization for standardization, 2012, 2017). The DNA extraction that followed the enrichment was performed with a commercial kit. Therefore, this method could be implemented more easily in routine, but it would still require a validation before accreditation. Metagenomics could even be performed in parallel to the research for an isolate with the routine method, as was done in our work on bacterial pathogens (Chapters 3-4), and give a result when no isolate can be found. Moreover, we showed in our comparison of Illumina and ONT on STEC spiked samples (Chapter 5) that a screening sequencing on Flongle flow cell, after extraction depleting the DNA from eukaryotic cells, could replace the conventional methods and possibly already give strain-level information and relatedness, at a low cost. This would offer opportunities to use this approach also for surveillance in the future (depending on available budgets from competent authorities), and not only for outbreak investigation. The data analysis still requires expertise as no push-button pipeline was developed yet (table 8.1), which has to be taken into account if implemented in routine. However, this could be developed in a follow-up project. Finally, we showed results on STEC and *Salmonella*, and this method could be transposed to other bacterial pathogens with minimal changes.

Our work on GMM contaminations (Chapter 6) demonstrated that we could obtain information on a sample for which no isolate could be cultured and for which no evidence that it was a genetically modified organism could be obtained with the set of routine tests available by using shotgun metagenomics. This could be done by using the same DNA extraction method as currently used at the NRL, which would facilitate the implementation in routine. More development is necessary in this field in order to improve the DNA extraction to obtain HMW DNA and automate the data analysis, which is still very complex and manual, but we showed the very strong potential of shotgun metagenomics to detect the GMM present in the sample in just one test and without *a priori* information. Sequencing with shotgun metagenomics might still represent a higher cost compared to the conventional methods (Table 8.1), but in the case of GMMs it might be the only method that can access the necessary information to prove that an unnatural construct is present in the sample, depending on the amount of tests that need to be done for full characterization and it replaces several time-consuming tests (Table 8.1). Therefore, this approach has a strong potential to be used in routine in the future, in particular when no isolate can be obtained but a suspicion of the presence of a GMM has arisen from the qPCR screening. Moreover, shotgun metagenomics potentially allows to obtain the entire genome of the microorganism, to compare it to other references from a database or develop new screening tests from the unnatural associations, and even to characterize several strains in the same sample. For this, the gathering of all known GMM sequences in database would greatly help the analysis (Figures 8.1 and 8.2).

Finally, our work on viral pathogens showed that not all contamination levels representative of real contaminations observed in routine might be detected. Nonetheless, we established that the use of a RNA extraction method more specific to viruses might improve these results as shown on our analysis of a previously published dataset. This is a very challenging case study as no enrichment was conducted and the contamination is low and heterogeneous. The sequencing at a higher sequencing depth could improve the results or limit of detection, but this would represent a higher cost which might not be affordable for routine tests. Ultimately, based on the results we presented (in Chapter 7) and other recent studies (Yang et al., 2017; Desdouits et al., 2020b), more work is necessary before metagenomics could be implemented for the detection and characterization of such food contaminants, although it is the only method that could provide a relatedness study for viral contaminants.

Because we worked with samples from the routine, following the current practices for sample preparation (Chapter 4 and 6) e.g. enrichment or not or used samples spiked at representative contamination doses (Chapters 3, 5 and 7) and could obtain a characterization of the contaminant comparable to the information obtained with conventional methods, we can conclude that the sensitivity of the method proved to be sufficient to comply with current routine practices. However, more tests are necessary to validate the method including determining precisely the parameters of the selected method such as sensitivity, specificity or limit of detection. In order to have statistically relevant results, a high number of samples should be tested (Negida et al., 2019; Peterson et al., 2022). This validation should also further address important aspects such as when should controls be included and which kind of controls, how to deal with mixed contaminations (e.g. mix of bacteria and viruses) and to determine precise levels to decide on the workflow to use. The bioinformatics analysis should also be validated and the standardized presentation of the data in a report for the competent authorities should also be determined. However, we believe that this PhD paved the way towards such validation by presenting the methods to be used or not (Figure 8.1). Finally, metagenomics has the potential to characterize multiple strains in a sample to the SNP level in one test, when sometimes not all strains are detected in routine. Using metagenomics would also allow to determine the occurrence of co-contaminations which is currently unknown or underestimated when using isolation.

8.7. Future challenges and perspectives

8.7.1. Future developments

Based on the answer given to these scientific questions, we have shown the potential of shotgun metagenomics to study microbiological contaminants in food to the strain level. For this, several workflows have been presented (Figure 8.1), where we have investigated the use of shotgun metagenomics for a specific case study, but representative for a typical scenario encountered by the NRL. However, the main challenge that still remains to be overcome is to circumvent the need for culture enrichment when working with a complex matrix (Figure 8.1), that we used to study bacterial pathogens (Chapters 3-4-5). Indeed, culturing adds about one day to the analysis, but above all, it is not a fully open approach: not all foodborne contaminants grow at 37°C in buffered peptone water in aerobic conditions (e.g. *Clostridium* (Edwards et al., 2013), *Campylobacter* (Davis and DiRita, 2008), GMMs with auxotrophy (Fraiture et al., 2020e)…). Adding culturing conditions would only multiply the number of samples to be analysed by metagenomics. Moreover, some competition during the culture might hinder the growth of the contaminant of interest (Heir et al., 2018; Ibrahim et al., 2021; Liu et al., 2021a). As showed in our work on GMM and norovirus, the contamination load in the sample has a great influence on the level of information that can be obtained with shotgun metagenomics without enrichment.

Alternative methods could allow to obtain strain-level information for very low levels of contaminations in complex matrices without culture enrichment (Figure 8.2). At short term, as discussed when studying norovirus, adaptive sampling could be considered as a way to enhance the sequencing ratio of the contaminant of interest. It is a relatively new tool proposed by ONT, allowing to continue or not to sequence a DNA fragment if it matches or not a database or reference genomes, *a priori* defined by the user. Indeed, the tool can be used to stop the sequencing of DNA fragments similar to the references in the database (depletion, used with the sequence genome of the food matrix, but these sequences might be computationally too heavy/large and not available for all matrices), or to stop sequencing any fragment that does not match the references in the database (enrichment, used with the sequence genomes of the contaminants of interest, ideally a database for all foodborne pathogens). As a result, the sequencing capacity will be devoted to the target of interest, and not to the background. Although this method was unsuccessful in our study of norovirus contamination (Chapter 7), the results could be improved if longer reads can be obtained. Indeed, several hundred base pairs have to be sequenced before the tool can decide to reject the DNA fragment or not (Martin et al., 2022). Moreover, the new ONT library preparation kit (kit 14) associated with the newest flow cells (R 10.4.1) should also allow to adjust the sequencing speed of the pores in the flow cell. This could increase the accuracy of the sequencing and therefore also impact the precision of the decision of the adaptive sampling, at the cost of a lower output (Brown, 2022). If the DNA concentration is too low for a sequencing, random DNA amplification could be considered as a possible way to obtain a higher level of genetic material instead of the replication by culture enrichment, as we showed that it does not negatively impact the result, but also did not improve it (Chapter 3), and ONT requires a high amount of input DNA.

Although we showed that metagenomics can characterize the contaminant and detect ARGs at strain level in food and feed samples, we were not yet able to determine with certitude if this gene was present on a plasmid or on the chromosome. This is particularly important as ARGs present on plasmids are known to be more easily transmissible in the environment (EFSA, 2021b). Some bioinformatics tools have been developed based on machine learning to evaluate if the ARG is present on a plasmid (Andreopoulos et al., 2022), but this is based on statistics and not on real information from the bacteria in our sample (Carattoli and Hasman, 2020), and we know that some regions of plasmids can be integrated in the genomes of the bacterium (Berbers et al., 2020). Moreover, the possibility for a false assumption would be even higher when studying artificial constructs in the case of GMMs. A new method allows to make the link between the plasmid and the bacterium holding it. Hi-C (Figure 8.2) labels spatially close portions of genomes (Lieberman-Aiden et al., 2009). It has been used to study structural configurations in the genomes (Schöpflin et al., 2022; Okabe and Kaneda, 2023) but also to link plasmid to the chromosome of the host bacterium (Cuscó et al., 2022; Kalmar et al., 2022). One obstacle, however, could be the low amount of intact cells in the food products as well as the complex data analysis. Another alternative method that has been described lately (mainly in clinical and environmental studies) could be tested as well i.e. single cell metagenomics (Figure 8.2). This method allows to sequence individual cells (obtained through various processes including microfluidics). Each single genome is then amplified and sequenced. Such method is of particular interest to attain strain level when multiple strains of the same species are present (Ide et al., 2022), or to link immediately a plasmid to its host at cell level (Nishikawa et al., 2022). It can also target specifically only the living or unaltered cells (Xu and Zhao, 2018; Chijiiwa et al., 2020). However, it can be biased due to the amplification step (Arikawa et al., 2021). Both of these approaches (i.e. Hi-C and single-cell metagenomics) are still very new and costly and have yet to be tested for the application of the contaminants in food or feed. It should be tested in order to determine how many cells have to be sequenced to obtain the information on the contaminants in the sample, and if this method is cost-effective (especially in terms of routine applicability) compared to other methods presented in this PhD.

In 2019, EFSA called for proofs of concepts of the use of shotgun metagenomics for the application of food safety and foodborne outbreak investigations (EFSA, 2019b). Our work presented such proofs of concepts on various matrices and contaminants, representative of realistic scenarios for routine laboratories. Therefore, the call of EFSA has been answered and the added value of shotgun metagenomics has been stated. The next step would include a thorough validation of these methods, for specific pathogens of interest, before applying them in a routine setting. This validation might imply some modification to the protocols previously presented (Figure 8.1). Moreover, user-friendly data analysis tools should be developed for ease of use in routine laboratories, e.g. through the implementation of the workflow as a webtool, and ultimately the competent authorities have to be convinced that shotgun metagenomics is a trusted alternative, and worthwile to pay for.

8.7.2. Entering the era of metagenomics

Although new developments are still necessary, the work presented in this PhD dissertation along with other recent publications (see examples in this section), has shown that metagenomics is able to profile microbial populations at strain level in one test. Several proofs of concepts have demonstrated that it is a fast and reliable method that can type a genome, and scientists have declared that we might now be entering a new era where metagenomics has proven that it can be used without the need for a systematic validation from isolates (Cocolin and Ercolini, 2015). This offers possibilities to use shotgun metagenomics for the characterization of foodborne contaminants but also for example for the diagnosis in clinical samples, allowing to obtain rapid information about the treatment to follow, or for rapid on site testing of pathogen or AMR presence to guide veterinaries in their treatment procedure or clinicians at locations where laboratories are not in the direct vicinity. Moreover, metagenomics also has the potential to study co-infections, which was done for example during the covid-19 pandemic (Molina-Mora et al., 2022), but this has only been described in very few studies.

The global COVID-19 pandemic that started in 2020 has increased the awareness on the emergence of new contaminants or the re-emergences and spread of known diseases (Morens and Fauci, 2020). For a good management of these epidemics, these contaminants should be detected, characterized and monitored, and data should be easily shared between countries. Shotgun metagenomics offers the possibility to obtain the data of an unknown novel pathogen, or gene (e.g. ARGs) without *a priori* information, from environmental or clinical samples, and therefore to detect it and characterize it (Miller et al., 2013; Govender, 2021). It has been rapidly shown that the occurrence of SARS-CoV-2 could be detected and monitored from the analysis of wastewater using qPCR (Lodder and de Roda Husman, 2020), but metagenomics also allows to explore the entire genetic material in these samples, permitting to possibly detect new variants or even new viruses, bacterial pathogens or ARGs (Hendriksen et al., 2019; Martínez-Puchol et al., 2021; Brumfield et al., 2022). The systematic surveillance of wastewater using shotgun metagenomics would endorse the gathering of all this epidemiological information at the level of cities or countries, and because bioinformatics allows retrospective analysis of previously sequenced metagenomes (Andersen et al., 2015), a trace back of novel contaminants is also conceivable.

Metagenomics has also now increasingly been used to study the human microbiomes, as a method to look at the whole microbial community and how it can impact the health of the patient, instead of looking for one contaminant. It has been used to study the diversity within the human gut, but also to detect ARGs within the genomes found in these communities (Qiu et al., 2020), and it is now becoming a tool of choice to make links between diseases and microbiome. Dysbiosis, or imbalance in the gut microbiota, has been associated with diabetes, obesity, allergies and cancers amongst others (Turnbaugh et al., 2009; Qin et al., 2012; Ahn et al., 2013; Le Chatelier et al., 2013; Abrahamsson et al., 2014; Wang et al., 2015). Dysbioses have also recently been linked to the consumption of xenobiotics in the diet (Chi et al., 2021; EFSA, 2022). Therefore, the study of the gut microbiomes can help determine new associations between illnesses and the food ingested. Determining the composition of the gut microbiomes with metagenomics can therefore play a role in food safety but also be used as a biomarker for diagnostic and therapeutic purposes. Finally, it has also been associated withhealth risk assessment to certain exposures or differential susceptibility to e.g. foodborne pathogens (Stevens et al., 2021; Strain et al., 2022) or SARS-CoV-2 (Sarkar et al., 2021; Veziant et al., 2021; Metwaly et al., 2022). Studies have now shown that the strain level is also necessary to accurately and precisely differentiate gut microbiomes, while metabarcoding cannot separate them (De Filippis et al., 2016, 2019). The focus has now also shifted to other parts of the human body such as skin, vagina or saliva (Nagar and Hasija, 2018; Liu et al., 2021b; Coker et al., 2022) but there is still much to uncover.

These new discoveries and application areas will have an important impact on the future of public health. Strain-level shotgun metagenomics methods such as these presented in this work will contribute to unravelling all the information from these microbiomes. The foundation stones have been laid towards fascinating new studies in the era of metagenomics.

References

Abrahamsson, T. R., Jakobsson, H. E., Andersson, A. F., Björkstén, B., Engstrand, L., and Jenmalm, M. C. (2014). Low gut microbiota diversity in early infancy precedes asthma at school age. Clin. Exp. Allergy 44, 842–850. doi:10.1111/cea.12253.

Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Ech, M., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 46, W537–W544. doi:10.1093/nar/gky379.

AFSCA (2019). Communiqué de presse de l'Agence régionale "Zorg en Gezondheid" et de l'Agence fédérale pour la Sécurité de la Chaîne alimentaire: Résultats de l'enquête sur le foyer de salmonelles à l'école hôtelière Spermalie à Bruges. Available at: http://www.afsca.be/professionnels/publications/presse/2019/2019-09-23b.asp.

Ahn, J., Sinha, R., Pei, Z., Dominianni, C., Wu, J., Shi, J., et al. (2013). Human Gut Microbiome and Risk for Colorectal Cancer. JNCI J. Natl. Cancer Inst. 105, 1907–1911. doi:10.1093/jnci/djt300.

Ahn, T. H., Chai, J., and Pan, C. (2015). Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance. Bioinformatics 31, 170–177. doi:10.1093/bioinformatics/btu641.

Andersen, L. O., Bonde, I., Nielsen, H. B., and Stensvold, C. R. (2015). A retrospective metagenomics approach to studying Blastocystis. FEMS Microbiol. Ecol. 91, 1–9. doi:10.1093/femsec/fiv072.

Andersen, S. C., Fachmann, M. S. R., Kiil, K., Nielsen, E. M., and Hoorfar, J. (2017). Gene-based pathogen detection: Can we use qPCR to predict the outcome of diagnostic metagenomics? Genes (Basel). 8. doi:10.3390/genes8110332.

Andreopoulos, W. B., Geller, A. M., Lucke, M., Balewski, J., Clum, A., Ivanova, N. N., et al. (2022). Deeplasmid: deep learning accurately separates plasmids from bacterial chromosomes. Nucleic Acids Res. 50, e17–e17. doi:10.1093/nar/gkab1115.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Appelt, S., Armougom, F., Bailly, M. Le, Robert, C., and Drancourt, M. (2014). Polyphasic analysis of a middle ages coprolite microbiota, Belgium. PLoS One 9, 1–8. doi:10.1371/journal.pone.0088376.

Arbeit, R. D., Arthur, M., Dunn, R., Kim, C., Selander, R. K., and Goldstein, R. (1990). Resolution of Recent Evolutionary Divergence among *Escherichia coli* from Related Lineages: The Application of Pulsed Field Electrophoresis to Molecular Epidemiology. J. Infect. Dis. 161, 230–235. doi:10.1093/infdis/161.2.230.

Arikawa, K., Ide, K., Kogawa, M., Saeki, T., Yoda, T., Endoh, T., et al. (2021). Recovery of strain-resolved genomes from human microbiome through an integration framework of single-cell genomics and metagenomics. Microbiome 9, 202. doi:10.1186/s40168-021-01152-4.

Avershina, E., Frye, S. A., Ali, J., Taxt, A. M., and Ahmad, R. (2022). Ultrafast and Cost-Effective Pathogen Identification and Resistance Gene Detection in a Clinical Setting Using Nanopore Flongle Sequencing. Front. Microbiol. 13. doi:10.3389/fmicb.2022.822402.

Aw, T. G., Wengert, S., and Rose, J. B. (2016). Metagenomic analysis of viruses associated with field-grown and retail lettuce identifies human and animal viruses. Int. J. Food Microbiol. 223, 50–56. doi:10.1016/j.ijfoodmicro.2016.02.008.

Baert, L., Mattison, K., Loisy-Hamon, F., Harlow, J., Martyres, A., Lebeau, B., et al. (2011). Review: Norovirus prevalence in Belgian, Canadian and French fresh produce: A threat to human health? Int. J. Food Microbiol. 151, 261–269. doi:10.1016/j.ijfoodmicro.2011.09.013.

Baker, C. A., Rubinelli, P. M., Park, S. H., Carbonero, F., and Ricke, S. C. (2016). Shiga toxin-producing *Escherichia coli* in food: Incidence, ecology, and detection strategies. Food Control 59, 407–419. doi:10.1016/j.foodcont.2015.06.011.

Bal, A., Pichon, M., Picard, C., Casalegno, J. S., Valette, M., Schuffenecker, I., et al. (2018). Quality control implementation for universal characterization of DNA and RNA viruses in clinical respiratory samples using single metagenomic next-generation sequencing workflow. bioRxiv, 1–10. doi:10.1101/367367.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477. doi:10.1089/cmb.2012.0021.

Barbau-Piednoir, E., Bertrand, S., Mahillon, J., Roosens, N. H., and Botteldoorn, N. (2013). SYBR®Green qPCR *Salmonella* detection system allowing discrimination at the genus, species and subspecies levels. Appl. Microbiol. Biotechnol. 97, 9811–9824. doi:10.1007/s00253-013-5234-x.

Barbau-piednoir, E., De Keersmaecker, S. C. J., Delvoye, M., Gau, C., Philipp, P., and Roosens, N. H. (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 Biotechnol. 15, 1–10. doi:10.1186/s12896-015-0216-y.

Barbau-Piednoir, E., de Keersmaecker, S. C. J., Wuyts, V., Gau, C., Pirovano, W., Costessi, A., et al. (2016). Genome sequence of EU-unauthorized genetically modified *Bacillus subtilis* strain 2014-3557 overproducing riboflavin, isolated from a vitamin B2 80% feed additive. Genome Announc. 3, 2014–2015. doi:10.1128/genomeA.00214-15.

Barbau-piednoir, E., Denayer, S., Botteldoorn, N., Dierick, K., Keersmaecker, S. C. J. De, and Roosens, N. H. (2018). Detection and discrimination in food samples of five *E. coli* pathotypes using a Combinatory SYBR®Green qPCR screening system. Applied Microbiology and Biotechnology 102, 3267–3285. ISBN: 0025301888.

Bartsch, C., Höper, D., Mäde, D., and Johne, R. (2018). Analysis of frozen strawberries involved in a large norovirus gastroenteritis outbreak using next generation sequencing and digital PCR. Food Microbiol. 76, 390–395. doi:10.1016/j.fm.2018.06.019.

Bavelaar, H. H. J., Rahamat-Langendoen, J., Niesters, H. G. M., Zoll, J., and Melchers, W. J. G. (2015). Whole genome sequencing of fecal samples as a tool for the diagnosis and genetic characterization of norovirus. J. Clin. Virol. 72, 122–125. doi:10.1016/j.jcv.2015.10.003.

Bentley, S. D., Chater, K. F., Cerdeño-Tárraga, A.-M., Challis, G. L., Thomson, N. R., James, K. D., et al. (2002). Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature 417, 141–147. doi:10.1038/417141a.

Berbers, B., Saltykova, A., Garcia-Graells, C., Philipp, P., Arella, F., Marchal, K., et al. (2020). Combining short and long read sequencing to characterize antimicrobial resistance genes on plasmids applied to an unauthorized genetically modified *Bacillus*. Sci. Rep. 10, 1–13. doi:10.1038/s41598-020-61158-0.

Bethesda (MD): National Library of Medicine (US), N. C. for B. I. (1988). National Center for Biotechnology Information (NCBI). https://www.ncbi.nlm.nih.gov/.

Bidawid, S., Farber, J. M., and Sattar, S. A. (2000). Contamination of Foods by Food Handlers: Experiments on Hepatitis A Virus Transfer to Food and Its Interruption. Appl. Environ. Microbiol. 66, 2759–2763. doi:10.1128/AEM.66.7.2759-2763.2000.

Bogaerts, B., Winand, R., Fu, Q., Van Braekel, J., Ceyssens, P. J., Mattheus, W., et al. (2019). Validation of a bioinformatics workflow for routine analysis of whole-genome sequencing data and related challenges for pathogen typing in a European national reference center: *Neisseria meningitidis* as a Proof-of-Concept. Front. Microbiol. 10. doi:10.3389/fmicb.2019.00362.

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

Bonnet, M., Lagier, J. C., Raoult, D., and Khelaifia, S. (2020). Bacterial culture through selective and non-selective conditions: the evolution of culture media in clinical microbiology. New Microbes New Infect. 34. doi:10.1016/j.nmni.2019.100622.

Bortolaia, V., Kaas, R. S., Ruppe, E., Roberts, M. C., Schwarz, S., Cattoir, V., et al. (2020). ResFinder 4.0 for predictions of phenotypes from genotypes. J. Antimicrob. Chemother. 75, 3491–3500. doi:10.1093/jac/dkaa345.

Bosch, A., Gkogka, E., Le, F. S., Loisy-hamon, F., Lee, A., Lieshout, L. Van, et al. (2020). Foodborne viruses: Detection, risk assessment, and control options in food processing. Int. J. Food Microbiol. 285, 110-128. doi:10.1016/j.ijfoodmicro.2018.06.001

Bottichio, L., Keaton, A., Thomas, D., Fulton, T., Tiffany, A., Frick, A., et al. (2020). Shiga Toxin–Producing *Escherichia coli* Infections Associated With Romaine Lettuce—United States, 2018. Clin. Infect. Dis. 71, e323–e330. doi:10.1093/cid/ciz1182.

Braeye, T., Denayer, S., De Rauw, K., Forier, A., Verluyten, J., Fourie, L., et al. (2014). Lessons learned from a textbook outbreak: EHEC-O157:H7 infections associated with the consumption of raw meat products, June 2012, Limburg, Belgium. Arch. Public Heal. 72, 44. doi:10.1186/2049-3258-72-44.

Brown, B., Allard, M., Bazaco, M. C., Blankenship, J., and Minor, T. (2021). An economic evaluation of the Whole Genome Sequencing source tracking program in the U.S. PLoS One 16, e0258262. doi:10.1371/journal.pone.0258262.

Brown, C. G. (2022). Oxford Nanopore technology update: CTO Clive G Brown unveils latest sequencing chemistry with highest performance to date, Short Fragment Mode and latest methylation performance evaluations. nanoporetech.com. Available at: https://nanoporetech.com/about-us/news/oxford-nanopore-technology-update-cto-clive-g-brown-unveils-latest-sequencing.

Brown, J. R., Roy, S., Ruis, C., Yara Romero, E., Shah, D., Williams, R., et al. (2016). Norovirus whole-genome sequencing by SureSelect target enrichment: A robust and sensitive method. J. Clin. Microbiol. 54, 2530–2537. doi:10.1128/JCM.01052-16.

Brumfield, K. D., Leddy, M., Usmani, M., Cotruvo, J. A., Tien, C., Dorsey, S., et al. (2022). Microbiome Analysis for Wastewater Surveillance during COVID-19. 13, 1–25. mBio, 13(4), e0059122. doi:10.1128/mbio.00591-22.

Brusa, V., Piñeyro, P. E., Galli, L., Linares, L. H., Ortega, E. E., Padola, N. L., et al. (2016). Isolation of Shiga toxin-producing *Escherichia coli* from ground beef using multiple combinations of enrichment broths and selective agars. Foodborne Pathog. Dis. 13, 163–170. doi:10.1089/fpd.2015.2034.

Butcher, H., Elson, R., Chattaway, M. A., Featherstone, C. A., Willis, C., Jorgensen, F., et al. (2016). Whole genome sequencing improved case ascertainment in an outbreak of Shiga toxin-producing *Escherichia coli* O157 associated with raw drinking milk. Epidemiol. Infect. 144, 2812–2823. doi:10.1017/S0950268816000509.

Buytaers, F. E., Fraiture, M.-A., Berbers, B., Vandermassen, E., Hoffman, S., Papazova, N., et al. (2021a). A shotgun metagenomics approach to detect and characterize unauthorized genetically modified microorganisms in microbial fermentation products. Food Chem. Mol. Sci. 2, 100023. doi:10.1016/j.fochms.2021.100023.

Buytaers, F. E., Saltykova, A., Denayer, S., Verhaegen, B., Vanneste, K., Roosens, N. H. C., et al. (2020). A Practical Method to Implement Strain-Level Metagenomics-Based Foodborne Outbreak Investigation and Source Tracking in Routine. Microorganisms 8, 1191. doi:10.3390/microorganisms8081191.

Buytaers, F. E., Saltykova, A., Denayer, S., Verhaegen, B., Vanneste, K., Roosens, N. H. C., et al. (2021b). Towards Real-Time and Affordable Strain-Level Metagenomics-Based Foodborne Outbreak Investigations Using Oxford Nanopore Sequencing Technologies. Front. Microbiol. 12, 1–13. doi:10.3389/fmicb.2021.738284.

Buytaers, F. E., Saltykova, A., Mattheus, W., Verhaegen, B., Roosens, N. H. C., Vanneste, K., et al. (2021c). Application of a strain-level shotgun metagenomics approach on food samples: resolution of the source of a *Salmonella* food-borne outbreak. Microb. Genomics 7. doi:10.1099/mgen.0.000547.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: Architecture and applications. BMC Bioinformatics 10, 1–9. doi:10.1186/1471-2105-10-421.

Carattoli, A., and Hasman, H. (2020). PlasmidFinder and In Silico pMLST: Identification and Typing of Plasmid Replicons in Whole-Genome Sequencing (WGS), Horizontal Gene Transfer. Methods in Molecular Biology, vol 2075, 285–294. doi:10.1007/978-1-4939-9877-7_20.

Carleton, H. A., Besser, J., Williams-Newkirk, A. J., Huang, A., Trees, E., and Gerner-Smidt, P. (2019). Metagenomic Approaches for Public Health Surveillance of Foodborne Infections: Opportunities and Challenges. Foodborne Pathog. Dis. 16, 474–479. doi:10.1089/fpd.2019.2636.

Carter, M. Q., Quinones, B., He, X., Zhong, W., Louie, J. W., Lee, B. G., et al. (2016). Clonal population exhibits high-level phenotypic variation that includes virulence traits. Appl. Enviromental Microbiol. 82, 1090–1101. doi:10.1128/AEM.03172-15.Editor.

Cassini, A., Högberg, L. D., Plachouras, D., Quattrocchi, A., Hoxha, A., Simonsen, G. S., et al. (2019). Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. Lancet Infect. Dis. 19, 56–66. doi:10.1016/S1473-3099(18)30605-4.

Centre National de Référence Salmonella & Shigella (2020). Rapport Annuel 2019. Available at: https://nrchm.wiv-isp.be/fr/centres_ref_labo/salmonella_et_shigella_spp/Rapports/Salmonella+Shigella 2019.pdf.

Charalampous, T., Kay, G. L., Richardson, H., Aydin, A., Baldan, R., Jeanes, C., et al. (2019). Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. Nat. Biotechnol. 37, 783–792. doi:10.1038/s41587-019-0156-5.

Chen, M. Y., Chen, W. C., Chen, P. C., Hsu, S. W., and Lo, Y. C. (2016). An outbreak of norovirus gastroenteritis associated with asymptomatic food handlers in Kinmen, Taiwan. BMC Public Health 16, 1–6. doi:10.1186/s12889-016-3046-5.

Cheung, M., Li, L., Nong, W., and Kwan, H. (2011). 2011 German *Escherichia coli* O104:H4 outbreak: Whole-genome phylogeny without alignment. BMC Res. Notes 4, 533. doi:10.1186/1756-0500-4-533.

Chi, L., Tu, P., Ru, H., and Lu, K. (2021). Studies of xenobiotic-induced gut microbiota dysbiosis: from correlation to mechanisms. Gut Microbes 13. doi:10.1080/19490976.2021.1921912.

Chijiiwa, R., Hosokawa, M., Kogawa, M., Nishikawa, Y., Ide, K., Sakanashi, C., et al. (2020). Single-cell genomics of uncultured bacteria reveals dietary fiber responders in the mouse gut microbiota. Microbiome 8, 5. doi:10.1186/s40168-019-0779-2.

Ching, C., Orubu, E. S. F., Sutradhar, I., Wirtz, V. J., Boucher, H. W., and Zaman, M. H. (2020). Bacterial antibiotic resistance development and mutagenesis following exposure to subinhibitory concentrations of fluoroquinolones in vitro : a systematic review of the literature. JAC-Antimicrobial Resist. 2. doi:10.1093/jacamr/dlaa068.

Cibulski, S., Alves de Lima, D., Fernandes dos Santos, H., Teixeira, T. F., Tochetto, C., Mayer, F. Q., et al. (2021). A plate of viruses: Viral metagenomics of supermarket chicken, pork and beef from Brazil. Virology 552, 1–9. doi:10.1016/j.virol.2020.09.005.

Cocolin, L., and Ercolini, D. (2015). Zooming into food-associated microbial consortia: A “cultural” evolution. Curr. Opin. Food Sci. 2, 43–50. doi:10.1016/j.cofs.2015.01.003.

Coker, M. O., Lebeaux, R. M., Hoen, A. G., Moroishi, Y., Gilbert-Diamond, D., Dade, E. F., et al. (2022). Metagenomic analysis reveals associations between salivary microbiota and body composition in early childhood. Sci. Rep. 12, 13075. doi:10.1038/s41598-022-14668-y.

Conceição-Neto, N., Zeller, M., Lefrère, H., De Bruyn, P., Beller, L., Deboutte, W., et al. (2015). Modular approach to customise sample preparation procedures for viral metagenomics: A reproducible protocol for virome analysis. Sci. Rep. 5, 1–14. doi:10.1038/srep16532.

Cosentino, S., Voldby Larsen, M., Møller Aarestrup, F., and Lund, O. (2013). PathogenFinder - Distinguishing Friend from Foe Using Bacterial Whole Genome Sequence Data. PLoS One 8. doi:10.1371/journal.pone.0077302.

Couto, N., Schuele, L., Raangs, E. C., Machado, M. P., Mendes, C. I., Jesus, T. F., et al. (2018). Critical steps in clinical shotgun metagenomics for the concomitant detection and typing of microbial pathogens. Sci. Rep. 8, 1–13. doi:10.1038/s41598-018-31873-w.

Cuscó, A., Pérez, D., Viñes, J., Fàbregas, N., and Francino, O. (2022). Novel canine high-quality metagenome-assembled genomes, prophages and host-associated plasmids provided by long-read metagenomics together with Hi-C proximity ligation. Microb. Genomics 8. doi:10.1099/mgen.0.000802.

D’aes, J., Fraiture, M.-A., Bogaerts, B., De Keersmaecker, S. C., Roosens, N. H. C., and Vanneste, K. (2022). Metagenomic characterization of multiple genetically modified *Bacillus* contaminations in commercial microbial fermentation products. Life 12(12),1971. doi: 10.3390/life12121971

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve years of SAMtools and BCFtools. Gigascience 10. doi:10.1093/gigascience/giab008.

Davies, A. R., Board, R. J., and Board, R. G. (1998). Microbiology of Meat and Poultry. Blackie Academic and Professional, London.

Davis, L., and DiRita, V. (2008). Growth and Laboratory Maintenance of *Campylobacter jejuni*. Curr. Protoc. Microbiol. 10. doi:10.1002/9780471729259.mc08a01s10.

De Bruyne, K., Slabbinck, B., Waegeman, W., Vauterin, P., De Baets, B., and Vandamme, P. (2011). Bacterial species identification from MALDI-TOF mass spectra through data analysis and machine learning. Syst. Appl. Microbiol. 34, 20–29. doi:10.1016/j.syapm.2010.11.003.

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., and Van Broeckhoven, C. (2018). NanoPack: Visualizing and processing long-read sequencing data. Bioinformatics 34, 2666–2669. doi:10.1093/bioinformatics/bty149.

De Filippis, F., Pasolli, E., Tett, A., Tarallo, S., Naccarati, A., De Angelis, M., et al. (2019). Distinct Genetic and Functional Traits of Human Intestinal *Prevotella copri* Strains Are Associated with Different Habitual Diets. Cell Host Microbe 25, 444-453.e3. doi:10.1016/j.chom.2019.01.004.

De Filippis, F., Pellegrini, N., Vannini, L., Jeffery, I. B., La Storia, A., Laghi, L., et al. (2016). High-level adherence to a Mediterranean diet beneficially impacts the gut microbiota and associated metabolome. Gut 65, 1812–1821. doi:10.1136/gutjnl-2015-309957.

De Keuckelaere, A., Li, D., Deliens, B., Stals, A., and Uyttendaele, M. (2015). Batch testing for noroviruses in frozen raspberries. Int. J. Food Microbiol. 192, 43–50. doi:10.1016/j.ijfoodmicro.2014.09.024.

De Rauw, K. ((PhD) S., Pierard, D. (Promotor), Scheutz, F. (Jury), Vandenberg, O. (Jury), Reynaert, H. (Jury), Hauser, B. (Jury), Crombe, F. (Jury), et al. (2019a). Detection, characterization and epidemiology of Shiga toxin-producing *Escherichia coli* in human infections.

De Rauw, K., Buyl, R., Jacquinet, S., and Piérard, D. (2019b). Risk determinants for the development of typical haemolytic uremic syndrome in Belgium and proposition of a new virulence typing algorithm for Shiga toxin-producing *Escherichia coli*. Epidemiol. Infect. 147, 0–4. doi:10.1017/S0950268818002546.

De Schrijver, K., Buvens, G., Possé, B., Van den Branden, D., Oosterlynck, O., De Zutter, L., et al. (2008). Outbreak of verocytotoxin-producing *E. coli* O145 and O26 infections associated with the consumption of ice cream produced at a farm, Belgium, 2007. Eurosurveillance 13, 9–10. doi:10.2807/ese.13.07.08041-en.

Deckers, M., Deforce, D., Fraiture, M. A., and Roosens, N. H. C. (2020a). Genetically modified micro-organisms for industrial food enzyme production: An overview. Foods 9. doi:10.3390/foods9030326.

Deckers, M., Vanneste, K., Winand, R., Hendrickx, M., Becker, P., De Keersmaecker, S. C. J., et al. (2020b). Screening strategy targeting the presence of food enzyme-producing fungi in food enzyme preparations. Food Control 117, 107295. doi:10.1016/j.foodcont.2020.107295.

Deckers, M., Vanneste, K., Winand, R., Keersmaecker, S. C. J. D., Denayer, S., Heyndrickx, M., et al. (2020c). Strategy for the identification of micro-organisms producing food and feed products: Bacteria producing food enzymes as study case. Food Chem. 305, 125431. doi:10.1016/j.foodchem.2019.125431.

Delahaye, C., and Nicolas, J. (2021). Sequencing DNA with nanopores: Troubles and biases. PLoS One 16. doi:10.1371/journal.pone.0257521.

Delaney, S., Murphy, R., and Walsh, F. (2018). A comparison of methods for the extraction of plasmids capable of conferring antibiotic resistance in a human pathogen from complex broiler cecal samples. Front. Microbiol. 9. doi:10.3389/fmicb.2018.01731.

Deng, X., den Bakker, H. C., and Hendriksen, R. S. (2016). Genomic Epidemiology: Whole-Genome-Sequencing–Powered Surveillance and Outbreak Investigation of Foodborne Bacterial Pathogens. Annu. Rev. Food Sci. Technol. 7, 353–374. doi:10.1146/annurev-food-041715-033259.

Desdouits, M., de Graaf, M., Strubbia, S., Oude Munnink, B. B., Kroneman, A., Le Guyader, F. S., et al. (2020a). Novel opportunities for NGS-based one health surveillance of foodborne viruses. One Heal. Outlook 2. doi:10.1186/s42522-020-00015-6.

Desdouits, M., Wacrenier, C., Ollivier, J., Schaeffer, J., and Le Guyader, F. S. (2020b). A Targeted Metagenomics Approach to Study the Diversity of Norovirus GII in Shellfish Implicated in Outbreaks. Viruses 12, 978. doi:10.3390/v12090978.

Dilthey, A. T., Jain, C., Koren, S., and Phillippy, A. M. (2019). Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. Nat. Commun. 10, 3066. doi:10.1038/s41467-019-10934-2.

Dubourg, G., Lamy, B., and Ruimy, R. (2018). Rapid phenotypic methods to improve the diagnosis of bacterial bloodstream infections: meeting the challenge to reduce the time to result. Clin. Microbiol. Infect. 24, 935–943. doi:10.1016/j.cmi.2018.03.031.

ECDC (2016). ECDC roadmap for integration of molecular and genomic typing into European-level surveillance and epidemic preparedness.

ECDC, and EFSA (2019). EFSA and ECDC technical report on the collection and analysis of whole genome sequencing data from food-borne pathogens and other relevant microorganisms isolated from human, animal, food, feed and food/feed environmental samples in the joint ECDC-EFSA mo. EFSA Support. Publ. 16. doi:10.2903/sp.efsa.2019.EN-1337.

ECDC, and EFSA (2020). JOINT ECDC-EFSA RAPID OUTBREAK ASSESSMENT: Multi-country outbreak of *Salmonella Enteritidis* infections linked to eggs, fourth update. doi:10.2903/sp.efsa.2020.EN-1799.

Eckert, S. E., Chan, J. Z. M., Houniet, D., Breuer, J., and Speight, G. (2016). Enrichment by hybridisation of long DNA fragments for Nanopore sequencing. Microb. genomics 2, e000087. doi:10.1099/mgen.0.000087.

Edwards, A. N., Suárez, J. M., and McBride, S. M. (2013). Culturing and Maintaining *Clostridium difficile* in an Anaerobic Environment. J. Vis. Exp. doi:10.3791/50787.

EFSA-ECDC (2022). The European Union Summary Report on Antimicrobial Resistance in zoonotic and indicator bacteria from humans, animals and food in 2019–2020. EFSA J. 20. doi:10.2903/j.efsa.2022.7209.

EFSA (2013). Scientific Opinion on the evaluation of molecular typing methods for major food-borne microbiological hazards and their use for attribution modelling, outbreak investigation and scanning surveillance: Part 1 (evaluation of methods and applications). EFSA J. 11, 1–84. doi:10.2903/j.efsa.2013.3502.

EFSA (2014a). Update of the technical specifications for harmonised reporting of food-borne outbreaks through the European Union reporting system in accordance with Directive 2003/99/EC. EFSA J. 12. doi:10.2903/j.efsa.2014.3598.

EFSA (2014b). Use of Whole Genome sequencing (WGS) of food-borne pathogens for public health protection. in doi:10.2805/66246.

EFSA (2016). Growth of spoilage bacteria during storage and transport of meat. EFSA J. 14. doi:10.2903/j.efsa.2016.4523.

EFSA (2018). The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2017. EFSA J. 16. doi:10.2903/j.efsa.2018.5500.

EFSA (2019a). The European Union One Health 2018 Zoonoses Report. doi:10.2903/j.efsa.2019.5926.

EFSA (2019b). Whole genome sequencing and metagenomics for outbreak investigation , source attribution and risk assessment of food-borne microorganisms. EFSA J. 17. doi:10.2903/j.efsa.2019.5898.

EFSA (2021a). EFSA statement on the requirements for whole genome sequence analysis of microorganisms intentionally used in the food chain. EFSA J., 1–13.

EFSA (2021b). Role played by the environment in the emergence and spread of antimicrobial resistance (AMR) through the food chain. EFSA J. 19. doi:10.2903/j.efsa.2021.6651.

EFSA (2021c). The European Union One Health 2019 Zoonoses Report. EFSA J. doi:10.2903/j.efsa.2021.6406.

EFSA (2021d). The European Union One Health 2020 Zoonoses Report. EFSA J. 19. doi:10.2903/j.efsa.2021.6971.

EFSA (2022). Microbiota analysis for risk assessment of xenobiotics: cumulative xenobiotic exposure and impact on human gut microbiota under One Health approach. EFSA J. 20. doi:10.2903/j.efsa.2022.e200916.

EFSA Panel on Biology Hazards (BIOHAZ) (2014). Scientific Opinion on the evaluation of molecular typing methods for major food-borne microbiological hazards and their use for attribution modelling, outbreak investigation and scanning surveillance: Part 2 (surveillance and data management activities). EFSA J. 12. doi:10.2903/j.efsa.2014.3784.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. Science. 323, 133–138. doi:10.1126/science.1162986.

Ellington, M. J., Ekelund, O., Aarestrup, F. M., Canton, R., Doumith, M., Giske, C., et al. (2017). The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. Clin. Microbiol. Infect. 23, 2–22. doi:10.1016/j.cmi.2016.11.012.

Escobar-Zepeda, A., De León, A. V. P., and Sanchez-Flores, A. (2015). The road to metagenomics: From microbiology to DNA sequencing technologies and bioinformatics. Front. Genet. 6, 1–15. doi:10.3389/fgene.2015.00348.

Ethelberg, S., Olsen, K. E. P., Scheutz, F., Jensen, C., Schiellerup, P., Engberg, J., et al. (2004). Virulence Factors for Hemolytic Uremic Syndrome, Denmark. Emerg. Infect. Dis. 10, 842–847. doi:10.3201/eid1005.030576.

European Commission (2011). Commision Staff Working Document: Lessons learned from the 2011 outbreak of Shiga toxin-producing *Escherichia coli* (STEC) O104:H4 in sprouted seeds. 22 p. Available at: http://ec.europa.eu/food/food/biosafety/salmonella/docs/cswd_lessons_learned_en.pdf.

European Parliament and the Concil of the European Union (2008). Regulation (EC) No 1333/2008. Off. J. Eur. Union.

European Parliament and the Council of the European Union (2003a). Regulation (EC) No 1829/2003. Off. J. Eur. Union. Available at: http://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX:32003R1829.

European Parliament and the Council of the European Union (2003b). REGULATION (EC) No 1830/2003. Off. J. Eur. Union.

European Parliament and the Council of the European Union (2008a). Regulation (EC) No 1331/2008. Off. J. Eur. Union.

European Parliament and the Council of the European Union (2008b). Regulation (EC) No 1332/2008. Off. J. Eur. Union.

FAO (2016). Applications of Whole Genome Sequencing in food safety management. Available at: http://www.fao.org/3/a-i5619e.pdf.

Feng, P. D., Weagant., S., and Jinneman, K. (2011). BAM: Diarrheagenic *Escherichia coli*. Accessed 30.07.19., 1–32. Available at: http://www.fda.gov/Food/FoodScienceResearch/LaboratoryMethods/ucm070080.htm.

Fernandez-Cassi, X., Timoneda, N., Gonzales-Gustavson, E., Abril, J. F., Bofill-Mas, S., and Girones, R. (2017). A metagenomic assessment of viral contamination on fresh parsley plants irrigated with fecally tainted river water. Int. J. Food Microbiol. 257, 80–90. doi:10.1016/j.ijfoodmicro.2017.06.001.

Fonager, J., Stegger, M., Rasmussen, L. D., Poulsen, M. W., Rønn, J., Andersen, P. S., et al. (2017). A universal primer-independent next-generation sequencing approach for investigations of norovirus outbreaks and novel variants. Sci. Rep. 7, 1–11. doi:10.1038/s41598-017-00926-x.

Food Safety Authority of Ireland (2019). Advice on Shiga toxin-producing *Escherichia coli* (STEC) detection in food. Available at: https://www.fsai.ie/publications/STEC_Report/

Fookes, M., Schroeder, G. N., Langridge, G. C., Blondel, C. J., Mammina, C., Connor, T. R., et al. (2011). *Salmonella bongori* Provides Insights into the Evolution of the Salmonellae. PLoS Pathog. 7, e1002191. doi:10.1371/journal.ppat.1002191.

Forbes, J. D., Knox, N. C., Peterson, C. L., and Reimer, A. R. (2018). Highlighting Clinical Metagenomics for Enhanced Diagnostic Decision-making: A Step Towards Wider Implementation. Comput. Struct. Biotechnol. J. 16, 108–120. doi:10.1016/j.csbj.2018.02.006.

Forbes, J. D., Knox, N. C., Ronholm, J., Pagotto, F., and Reimer, A. (2017). Metagenomics: The next culture-independent game changer. Front. Microbiol. 8, 1–21. doi:10.3389/fmicb.2017.01069.

Forghani, F., Li, S., Zhang, S., Mann, D. A., Deng, X., den Bakker, H. C., et al. (2020). *Salmonella enterica* and *Escherichia coli* in Wheat Flour: Detection and Serotyping by a Quasimetagenomic Approach Assisted by Magnetic Capture, Multiple-Displacement Amplification, and Real-Time Sequencing. Appl. Environ. Microbiol. 86, 1–15. doi:10.1128/AEM.00097-20.

Forsythe, S. J. (2000). The microbiology of safe food. Blackwell Science. doi:10.1002/jsfa.995.

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

Fraiture, M.-A., Gobbo, A., Papazova, N., and 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, 78. doi:10.3390/fermentation8020078.

Fraiture, M.-A., Marchesi, U., Verginelli, D., Papazova, N., and Roosens, N. H. C. (2021b). Development of a Real-time PCR Method Targeting an Unauthorized Genetically Modified Microorganism Producing Alpha-Amylase. Food Anal. Methods 14, 2211–2220. doi:10.1007/s12161-021-02044-x.

Fraiture, M. A., Bogaerts, B., Winand, R., Deckers, M., Papazova, N., Vanneste, K., et al. (2020a). Identification of an unauthorized genetically modified bacteria in food enzyme through whole-genome sequencing. Sci. Rep. 10, 1–12. doi:10.1038/s41598-020-63987-5.

Fraiture, M. A., Deckers, M., Papazova, N., and Roosens, N. H. C. (2020b). Are antimicrobial resistance genes key targets to detect genetically modified microorganisms in fermentation products? Int. J. Food Microbiol. 331, 108749. doi:10.1016/j.ijfoodmicro.2020.108749.

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

Fraiture, M. A., Deckers, M., Papazova, N., and Roosens, N. H. C. (2020d). Strategy to Detect Genetically Modified Bacteria Carrying Tetracycline Resistance Gene in Fermentation Products. Food Anal. Methods. doi:10.1007/s12161-020-01803-6.

Fraiture, M. A., Herman, P., Papazova, N., De Loose, M., Deforce, D., Ruttink, T., et al. (2017). An integrated strategy combining DNA walking and NGS to detect GMOs. Food Chem. 232, 351–358. doi:10.1016/j.foodchem.2017.03.067.

Fraiture, M. A., Herman, P., Papazova, N., and Roosens, N. H. (2015). UGM characterization using DNA walking strategy. LabInfo 14, 32–34. Available at: http://www.afsca.be/laboratories/labinfo/_documents/2015-12_labinfo14_en.pdf.

Fraiture, M. A., Joly, L., Vandermassen, E., Delvoye, M., Van Geel, D., Michelet, J. Y., et al. (2021c). Retrospective survey of unauthorized genetically modified bacteria harbouring antimicrobial resistance genes in feed additive vitamin B2 commercialized in Belgium: Challenges and solutions. Food Control 119, 107476. doi:10.1016/j.foodcont.2020.107476.

Fraiture, M. A., Papazova, N., and Roosens, N. H. C. (2020e). DNA walking strategy to identify unauthorized genetically modified bacteria in microbial fermentation products. Int. J. Food Microbiol. 337, 108913. doi:10.1016/j.ijfoodmicro.2020.108913.

Franz, E., Gras, L. M., and Dallman, T. (2016). Significance of whole genome sequencing for surveillance, source attribution and microbial risk assessment of foodborne pathogens. Curr. Opin. Food Sci. 8, 74–79. doi:10.1016/j.cofs.2016.04.004.

Fratamico, P. M., DebRoy, C., and Needleman, D. S. (2016). Editorial: Emerging Approaches for Typing, Detection, Characterization, and Traceback of *Escherichia coli*. Front. Microbiol. 7. doi:10.3389/fmicb.2016.02089.

Gand, M. (2020). Development of a genoserotyping system for the identification of Salmonella serotypes. Universiteit Gent. Faculteit Wetenschappen.

Gardy, J. L., and Loman, N. J. (2018). Towards a genomics-informed, real-time, global pathogen surveillance system. Nat. Rev. Genet. 19, 9–20. doi:10.1038/nrg.2017.88.

Govender, K. N. (2021). Precision pandemic preparedness: Improving diagnostics with metagenomics. J. Clin. Microbiol. 59, 1–6. doi:10.1128/JCM.02146-20.

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

Grädel, C., Angel Terrazos Miani, M., Barbani, M. T., Leib, S. L., Franziska, S.-R., and Ramette, A. (2019). Rapid and Cost-Efficient Enterovirus Genotyping from Clinical Samples Using Flongle Flow Cells. Genes (Basel). 10. doi: 10.3390/genes10090659

Greig, D. R., Jenkins, C., Gharbia, S., and Dallman, T. J. (2019). Comparison of single-nucleotide variants identified by Illumina and Oxford Nanopore technologies in the context of a potential outbreak of Shiga toxin–producing *Escherichia coli*. Gigascience 8. doi:10.1093/gigascience/giz104.

Greninger, A. L., Naccache, S. N., Federman, S., Yu, G., Mbala, P., Bres, V., et al. (2015). Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Med. 7, 1–13. doi:10.1186/s13073-015-0220-9.

Grützke, J., Gwida, M., Deneke, C., Brendebach, H., Projahn, M., Schattschneider, A., et al. (2021). Direct identification and molecular characterization of zoonotic hazards in raw milk by metagenomics using *brucella* as a model pathogen. Microb. Genomics 7. doi:10.1099/MGEN.0.000552.

Grützke, J., Malorny, B., Hammerl, J. A., Busch, A., Tausch, S. H., Tomaso, H., et al. (2019). Fishing in the Soup – Pathogen Detection in Food Safety Using Metabarcoding and Metagenomic Sequencing. 10, 1–15. doi:10.3389/fmicb.2019.01805.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32, 2847–2849. doi:10.1093/bioinformatics/btw313.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. Bioinformatics 29, 1072–1075. doi:10.1093/bioinformatics/btt086.

Halliday, M. L., Kang, L.-Y., Zhou, T.-K., Hu, M.-D., Pan, Q.-C., Fu, T.-Y., et al. (1991). An Epidemic of Hepatitis A Attributable to the Ingestion of Raw Clams in Shanghai, China. J. Infect. Dis. 164, 852–859. doi:10.1093/infdis/164.5.852.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. Chem. Biol. 5. doi:10.1016/S1074-5521(98)90108-9.

Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., et al. (2001). Complete genome sequence of enterohemorrhagic *Eschelichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res. 8, 11–22. doi:10.1093/dnares/8.1.11.

Heather, J. M., and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics 107, 1–8. doi:10.1016/j.ygeno.2015.11.003.

Heir, E., Møretrø, T., Simensen, A., and Langsrud, S. (2018). *Listeria monocytogenes* strains show large variations in competitive growth in mixed culture biofilms and suspensions with bacteria from food processing environments. Int. J. Food Microbiol. 275, 46–55. doi:10.1016/j.ijfoodmicro.2018.03.026.

Hendriksen, R. S., Munk, P., Njage, P., van Bunnik, B., McNally, L., Lukjancenko, O., et al. (2019). Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. Nat. Commun. 10. doi:10.1038/s41467-019-08853-3.

Hernando-Amado, S., Coque, T. M., Baquero, F., and Martínez, J. L. (2019). Defining and combating antibiotic resistance from One Health and Global Health perspectives. Nat. Microbiol. 4, 1432–1442. doi:10.1038/s41564-019-0503-9.

Hibbs, A. C., Secor, W. E., Van Gerven, D., and Armelagos, G. (2011). Irrigation and infection: The immunoepidemiology of schistosomiasis in ancient Nubia. Am. J. Phys. Anthropol. 145, 290–298. doi:10.1002/ajpa.21493.

Hill, A. A., Crotta, M., Wall, B., Good, L., O'Brien, S. J., and Guitian, J. (2017). Towards an integrated food safety surveillance system: A simulation study to explore the potential of combining genomic and epidemiological metadata. R. Soc. Open Sci. 4. doi:10.1098/rsos.160721.

Höper, D., Mettenleiter, T. C., and Beer, M. (2016). Metagenomic approaches to identifying infectious agents. Rev. Sci. Tech. 35, 83–93. doi:10.20506/rst.35.1.2419.

Huang, A. D., Luo, C., Pena-Gonzalez, A., Weigand, M. R., Tarr, C. L., and Konstantinidis, K. T. (2017). Metagenomics of two severe foodborne outbreaks provides diagnostic signatures and signs of coinfection not attainable by traditional methods. Appl. Environ. Microbiol. 83, 1–14. doi:10.1128/AEM.02577-16.

Hyeon, J., Li, S., Mann, D. A., Zhang, S., Li, Z., Chen, Y., et al. (2018). Quasimetagenomics-based and real-time-sequencing-aided detection and subtyping of *Salmonella enterica* from food samples. Appl. Environ. Microbiol. 84, 1–15. doi:10.1128/AEM.02340-17.

Ibrahim, S. A., Ayivi, R. D., Zimmerman, T., Siddiqui, S. A., Altemimi, A. B., Fidan, H., et al. (2021). Lactic Acid Bacteria as Antimicrobial Agents: Food Safety and Microbial Food Spoilage Prevention. Foods 10, 3131. doi:10.3390/foods10123131.

Ide, K., Saeki, T., Arikawa, K., Yoda, T., Endoh, T., Matsuhashi, A., et al. (2022). Exploring strain diversity of dominant human skin bacterial species using single-cell genome sequencing. Front. Microbiol. 13. doi:10.3389/fmicb.2022.955404.

Illumina (2022). High-performance long-read assay enables contiguous data with N50 of 6–7 kb on existing Illumina platforms. Available at: https://www.illumina.com/science/genomics-research/articles/infinity-high-performance-long-read-assay.html [Accessed November 4, 2022].

Inns, T., Ashton, P. M., Herrera-Leon, S., Lighthill, J., Foulkes, S., Jombart, T., et al. (2017). Prospective use of whole genome sequencing (WGS) detected a multi-country outbreak of *Salmonella* Enteritidis. Epidemiol. Infect. 145, 289–298. doi:10.1017/S0950268816001941.

Inouye, M., Dashnow, H., Raven, L. A., Schultz, M. B., Pope, B. J., Tomita, T., et al. (2014). SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Med. 6, 1–16. doi:10.1186/s13073-014-0090-6.

ISO: International Organization for standardization ISO/DIS 23418 Microbiology of the food chain — Whole genome sequencing for typing and genomic characterization of foodborne bacteria — General requirements and guidance.

ISO: International Organization for standardization (2012). ISO/TS 13136:2012 microbiology of food and animal feed - real-time polymerase chain reaction (PCR)-based method for the detetion of food-borne pathogens - Horizontal method for the detection of Shiga toxin-producing *Escherichia coli* (STEC) and the determination of O157, O111, O26, O103 and O145 serogroups.

ISO: International Organization for standardization (2017). ISO 6579-1:2017 Microbiology of the food chain — Horizontal method for the detection, enumeration and serotyping of Salmonella — Part 1: Detection of Salmonella spp.

ISO: International Organization for standardization (2019). ISO 15216-2:2019 Microbiology of the food chain — Horizontal method for determination of hepatitis A virus and norovirus using real-time RT-PCR.

Jain, S., Mukhopadhyay, K., and Thomassin, P. J. (2019). An economic analysis of *salmonella* detection in fresh produce, poultry, and eggs using whole genome sequencing technology in Canada. Food Res. Int. 116, 802–809. doi:10.1016/j.foodres.2018.09.014.

Jajarmi, M., Imani Fooladi, A. A., Badouei, M. A., and Ahmadi, A. (2017). Virulence genes, Shiga toxin subtypes, major O-serogroups, and phylogenetic background of Shiga toxin-producing *Escherichia coli* strains isolated from cattle in Iran. Microb. Pathog. 109, 274–279. doi:10.1016/j.micpath.2017.05.041.

Jasson, V., Jacxsens, L., Luning, P., Rajkovic, A., and Uyttendaele, M. (2010). Alternative microbial methods: An overview and selection criteria. Food Microbiol. 27, 710–730. doi:10.1016/j.fm.2010.04.008.

Jasson, V., Rajkovic, A., Baert, L., Debevere, J., and Uyttendaele, M. (2009). Comparison of enrichment conditions for rapid detection of low numbers of sublethally injured *Escherichia coli* 0157 in food. J. Food Prot. 72, 1862–1868. doi:10.4315/0362-028X-72.9.1862.

Jeong, S.-H., and Lee, H.-S. (2010). Hepatitis A: Clinical Manifestations and Management. Intervirology 53, 15–19. doi:10.1159/000252779.

Joensen, K. G., Scheutz, F., Lund, O., Hasman, H., Kaas, R. S., Nielsen, E. M., et al. (2014). Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. J. Clin. Microbiol. 52, 1501–1510. doi:10.1128/JCM.03617-13.

Joensen, K. G., Tetzschner, A. M. M., Iguchi, A., Aarestrup, F. M., and Scheutz, F. (2015). Rapid and easy in silico serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data. J. Clin. Microbiol. 53, 2410–2426. doi:10.1128/JCM.00008-15.

Jones, M. K., Grau, K. R., Costantini, V., Kolawole, A. O., de Graaf, M., Freiden, P., et al. (2015). Human norovirus culture in B cells. Nat. Protoc. 10, 1939–1947. doi:10.1038/nprot.2015.121.

Josefsen, M. H., Andersen, S. C., Christensen, J., and Hoorfar, J. (2015). Microbial food safety: Potential of DNA extraction methods for use in diagnostic metagenomics. J. Microbiol. Methods 114, 30–34. doi:10.1016/j.mimet.2015.04.016.

Juul, S., Izquierdo, F., Hurst, A., Dai, X., Wright, A., Kulesha, E., et al. (2015). What's in my pot? Real-time species identification on the MinION. bioRxiv, 030742. doi:10.1101/030742.

Kalmar, L., Gupta, S., Kean, I. R. L., Ba, X., Hadjirin, N., Lay, E. M., et al. (2022). HAM-ART: An optimised culture-free Hi-C metagenomics pipeline for tracking antimicrobial resistance genes in complex microbial communities. PLOS Genet. 18, e1009776. doi:10.1371/journal.pgen.1009776.

Kawai, T., Sekizuka, T., Yahata, Y., Kuroda, M., Kumeda, Y., Iijima, Y., et al. (2012). Identification *of Kudoa septempunctata* as the causative agent of novel food poisoning outbreaks in Japan by consumption of *paralichthys olivaceus* in raw fish. Clin. Infect. Dis. 54, 1046–1052. doi:10.1093/cid/cir1040.

Keim, P., Price, L. B., Klevytska, A. M., Smith, K. L., Schupp, J. M., Okinaka, R., et al. (2000). Multiple-Locus Variable-Number Tandem Repeat Analysis Reveals Genetic Relationships within *Bacillus anthracis*. J. Bacteriol. 182, 2928–2936. doi:10.1128/JB.182.10.2928-2936.2000.

Kinnula, S., Hemminki, K., Kotilainen, H., Ruotsalainen, E., Tarkka, E., Salmenlinna, S., et al. (2018). Outbreak of multiple strains of non-o157 shiga toxin-producing and enteropathogenic *escherichia coli* associated with rocket salad, Finland, autumn 2016. Eurosurveillance 23, 1–8. doi:10.2807/1560-7917.ES.2018.23.35.1700666.

Kircher, M., Sawyer, S., and Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. 40, 1–8. doi:10.1093/nar/gkr771.

Kleinheinz, K. A., Joensen, K. G., and Larsen, M. V. (2014). Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and *E.*

coli virulence genes in bacteriophage and prophage nucleotide sequences. Bacteriophage. 4(1):e27943. doi: 10.4161/bact.27943.

Kleppe, K., Ohtsuka, E., Kleppe, R., Molineux, I., and Khorana, H. G. (1971). Studies on polynucleotides. XCVI. Repair replications of short synthetic DNA's as catalyzed by DNA polymerases. J. Mol. Biol. 56, 341–361. doi:10.1016/0022-2836(71)90469-4.

Klümper, U., Riber, L., Dechesne, A., Sannazzarro, A., Hansen, L. H., Sørensen, S. J., et al. (2015). Broad host range plasmids can invade an unexpectedly diverse fraction of a soil bacterial community. ISME J. 9, 934–945. doi:10.1038/ismej.2014.191.

Knudsen, B. E., Bergmark, L., Munk, P., Lukjancenko, O., Priemé, A., Aarestrup, F. M., et al. (2016). Impact of Sample Type and DNA Isolation Procedure on Genomic Inference of Microbiome Composition. mSystems 1, e00095-16. doi:10.1128/mSystems.00095-16.

Kono, N., and Arakawa, K. (2019). Nanopore sequencing: Review of potential applications in functional genomics. Dev. Growth Differ. 61, 316–326. doi:10.1111/dgd.12608.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Res. 27, 722–736. doi:10.1101/gr.215087.116.

Kornberg, A., Lehman, I. R., Bessman, M. J., and Simms, E. S. (1956). Enzymic synthesis of deoxyribonucleic acid. Biochim. Biophys. Acta 21, 197–198. doi:10.1016/0006-3002(56)90127-5.

Koutsoumanis, K., Allende, A., Alvarez-Ordóñez, A., Bover-Cid, S., Chemaly, M., Davies, R., et al. (2020). Pathogenicity assessment of Shiga toxin-producing *Escherichia coli* (STEC) and the public health risk posed by contamination of food with STEC. EFSA J. 18. doi:10.2903/j.efsa.2020.5967.

Kovac, J., Bakker, H. den, Carroll, L. M., and Wiedmann, M. (2017). Precision food safety: A systems approach to food safety facilitated by genomics tools. TrAC - Trends Anal. Chem. 96, 52–61. doi:10.1016/j.trac.2017.06.001.

Kralik, P., and Ricchi, M. (2017). A basic guide to real time PCR in microbial diagnostics: Definitions, parameters, and everything. Front. Microbiol. 8, 1–9. doi:10.3389/fmicb.2017.00108.

Kroneman, A., Vennema, H., Deforche, K., Avoort, H. v. d., Peñaranda, S., Oberste, M. S., et al. (2011). An automated genotyping tool for enteroviruses and noroviruses. J. Clin. Virol. 51, 121–125. doi:10.1016/j.jcv.2011.03.006.

Krüger, A., Lucchesi, P. M. A., and Parma, A. E. (2011). Verotoxins in bovine and meat verotoxin-producing *Escherichia coli* isolates: Type, number of variants, and relationship to cytotoxicity. Appl. Environ. Microbiol. 77, 73–79. doi:10.1128/AEM.01445-10.

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. 35, 1547–1549. doi:10.1093/molbev/msy096.

Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., et al. (2013). Richness of human gut microbiome correlates with metabolic markers. Nature 500, 541–546. doi:10.1038/nature12506.

Le Minor, L. (1988). Typing of *Salmonella* species. Eur. J. Clin. Microbiol. Infect. Dis. 7, 214–218. doi:10.1007/BF01963091.

Lee, H., and Yoon, Y. (2021). Etiological agents implicated in foodborne illness world wide. Food Sci. Anim. Resour. 41, 1–7. doi:10.5851/KOSFA.2020.E75.

Leggett, R. M., Alcon-Giner, C., Heavens, D., Caim, S., Brook, T. C., Kujawska, M., et al. (2020). Rapid MinION profiling of preterm microbiota and antimicrobial-resistant pathogens. Nat. Microbiol. 5, 430–442. doi:10.1038/s41564-019-0626-z.

Leonard, S. R., Mammel, M. K., Lacher, D. W., and Elkins, C. A. (2015). Application of metagenomic sequencing to food safety: Detection of shiga toxin-producing *Escherichia coli* on fresh bagged spinach. Appl. Environ. Microbiol. 81, 8183–8191. doi:10.1128/AEM.02601-15.

Leonard, S. R., Mammel, M. K., Lacher, D. W., and Elkins, C. A. (2016). Strain-level discrimination of shiga toxin-producing *Escherichia coli* in spinach using metagenomic sequencing. PLoS One 11, 1–21. doi:10.1371/journal.pone.0167870.

Letunic, I., and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 44, W242–W245. doi:10.1093/nar/gkw290.

Lewandowski, K., Xu, Y., Pullan, S. T., Lumley, S. F., Foster, D., Sanderson, N., et al. (2019). Metagenomic Nanopore sequencing of influenza virus direct from clinical respiratory samples. bioRxiv 58, 1–15. doi:10.1101/676155.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31, 1674–1676. doi:10.1093/bioinformatics/btv033.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. doi:https://doi.org/10.48550/arXiv.1303.3997.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26, 589–595. doi:10.1093/bioinformatics/btp698.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. doi:10.1093/bioinformatics/btp352.

Li, Y., He, X. zhou, Li, M. hui, Li, B., Yang, M. jie, Xie, Y., et al. (2020). Comparison of third-generation sequencing approaches to identify viral pathogens under public health emergency conditions. Virus Genes 56, 288–297. doi:10.1007/s11262-020-01746-4.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 326, 289–293. doi:10.1126/science.1181369.

Liefting, L. W., Waite, D. W., and Thompson, J. R. (2021). Application of Oxford Nanopore Technology to Plant Virus Detection. Viruses 13, 1424. doi:10.3390/v13081424.

Liu, D., Walcott, R., Mis Solval, K., and Chen, J. (2021a). Influence of Bacterial Competitors on *Salmonella enterica* and Enterohemorrhagic *Escherichia coli* Growth in Microbiological Media and Attachment to Vegetable Seeds. Foods 10, 285. doi:10.3390/foods10020285.

Liu, D., Zhang, Z., Li, S., Wu, Q., Tian, P., Zhang, Z., et al. (2020). Fingerprinting of human noroviruses co-infections in a possible foodborne outbreak by metagenomics. Int. J. Food Microbiol. 333, 108787. doi:10.1016/j.ijfoodmicro.2020.108787.

Liu, F., Zhou, Y., Zhu, L., Wang, Z., Ma, L., He, Y., et al. (2021b). Comparative metagenomic analysis of the vaginal microbiome in healthy women. Synth. Syst. Biotechnol. 6, 77–84. doi:10.1016/j.synbio.2021.04.002.

Lodder, W., and de Roda Husman, A. M. (2020). SARS-CoV-2 in wastewater: potential health risk, but also data source. Lancet Gastroenterol. Hepatol. 5, 533–534. doi:10.1016/S2468-1253(20)30087-X.

Loman, N. J., Constantinidou, C., Chan, J. Z. M., Halachev, M., Sergeant, M., Penn, C. W., et al. (2012). High-throughput bacterial genome sequencing: An embarrassment of choice, a world of opportunity. Nat. Rev. Microbiol. 10, 599–606. doi:10.1038/nrmicro2850.

Loman, N. J., Constantinidou, C., Christner, M., Rohde, H., Chan, J. Z.-M., Quick, J., et al. (2013). A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. Jama 309, 1502–10. doi:10.1001/jama.2013.3231.

Loman, N. J., Quick, J., and Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat. Methods 12, 733–735. doi:10.1038/nmeth.3444.

Loong, S. K., Khor, C. S., Jafar, F. L., and AbuBakar, S. (2016). Utility of 16S rDNA Sequencing for Identification of Rare Pathogenic Bacteria. J. Clin. Lab. Anal. 30, 1056–1060. doi:10.1002/jcla.21980.

Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., et al. (1998). Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. 95, 3140–3145. doi:10.1073/pnas.95.6.3140.

Marcelino, V. R., Clausen, P. T. L. C., Buchmann, J. P., Wille, M., Iredell, J. R., Meyer, W., et al. (2020a). CCMetagen: Comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. Genome Biol. 21, 1–15. doi:10.1186/s13059-020-02014-2.

Marcelino, V. R., Holmes, E. C., and Sorrell, T. C. (2020b). The use of taxon-specific reference databases compromises metagenomic classification. BMC Genomics 21, 1–5. doi:10.1186/s12864-020-6592-2.

Marquet, M., Zöllkau, J., Pastuschek, J., Viehweger, A., Schleußner, E., Makarewicz, O., et al. (2022). Evaluation of microbiome enrichment and host DNA depletion in human vaginal samples using Oxford Nanopore's adaptive sequencing. Sci. Rep. 12, 1–10. doi:10.1038/s41598-022-08003-8.

Martin, S., Heavens, D., Lan, Y., Horsfield, S., Clark, M. D., and Leggett, R. M. (2022). Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples. Genome Biol. 23, 1–27. doi:10.1186/s13059-021-02582-x.

Martínez-Puchol, S., Itarte, M., Rusiñol, M., Forés, E., Mejías-Molina, C., Andrés, C., et al. (2021). Exploring the diversity of coronavirus in sewage during COVID-19 pandemic: Don't miss the forest for the trees. Sci. Total Environ. 800, 149562. doi:10.1016/j.scitotenv.2021.149562.

Mathijs, E., Denayer, S., Palmeira, L., Botteldoorn, N., Scipioni, A., Vanderplasschen, A., et al. (2011). Novel norovirus recombinants and GII.4 sub-lineages associated with outbreaks between 2006 and 2010 in Belgium. Virol. J. 8, 310. doi:10.1186/1743-422X-8-310.

Mcarthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., et al. (2013). The Comprehensive Antibiotic Resistance Database. Antimicrob Agents Chemother 57, 3348–3357. doi:10.1128/AAC.00419-13.

McMeekin, T. A. (2003). Detecting pathogens in food. Woodhead Publishing Limited.

Mestan, K. K., Ilkhanoff, L., Mouli, S., and Lin, S. (2011). Genomic sequencing in clinical trials. J. Transl. Med. 9, 222. doi:10.1186/1479-5876-9-222.

Metwaly, A., Reitmeier, S., and Haller, D. (2022). Microbiome risk profiles as biomarkers for inflammatory and metabolic disorders. Nat. Rev. Gastroenterol. Hepatol. 19, 383–397. doi:10.1038/s41575-022-00581-2.

Miller, R. R., Montoya, V., Gardy, J. L., Patrick, D. M., and Tang, P. (2013). Metagenomics for pathogen detection in public health. Genome Med. 5. doi:10.1186/gm485.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., et al. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. 37, 1530–1534. doi:10.1093/molbev/msaa015.

Molina-Mora, J. A., Cordero-Laurent, E., Calderón-Osorno, M., Chacón-Ramírez, E., and Duarte-Martínez, F. (2022). Metagenomic pipeline for identifying co-infections among distinct SARS-CoV-2 variants of concern: study cases from Alpha to Omicron. Sci. Rep. 12, 9377. doi:10.1038/s41598-022-13113-4.

Morens, D. M., and Fauci, A. S. (2020). Emerging Pandemic Diseases: How We Got to COVID-19. Cell 183, 837. doi:10.1016/j.cell.2020.10.022.

Morgan, M., Watts, V., Allen, D., Curtis, D., Kirolos, A., Macdonald, N., et al. (2019). Challenges of investigating a large food-borne norovirus outbreak across all branches of a restaurant group in the United Kingdom, October 2016. Eurosurveillance 24. doi:10.2807/1560-7917.ES.2019.24.18.1800511.

Murray, C. J., Ikuta, K. S., Sharara, F., Swetschinski, L., Robles Aguilar, G., Gray, A., et al. (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet 399, 629–655. doi:10.1016/S0140-6736(21)02724-0.

Nadon, C., Walle, I. Van, Gerner-smidt, P., Campos, J., Chinen, I., and Concepcion-acevedo, J. (2017). PulseNet International : Vision for the implementation of whole genome sequencing (WGS) for global food- borne disease surveillance. Euro Surveill. doi:10.2807/1560-7917.ES.2017.22.23.30544.

Nagar, P., and Hasija, Y. (2018). Metagenomic approach in study and treatment of various skin diseases: a brief review. Biomed. Dermatology 2, 19. doi:10.1186/s41702-018-0029-4.

Nagarajan, V., Chen, J.-S., Hsu, G.-J., Chen, H.-P., Chao, H.-C., Huang, S.-W., et al. (2022). Surveillance of Adenovirus and Norovirus Contaminants in the Water and Shellfish of Major Oyster Breeding Farms and Fishing Ports in Taiwan. Pathogens 11, 316. doi:10.3390/pathogens11030316.

Nanopore Protocol (2019). Genomic DNA by Ligation (SQK-LSK-109) version GDE_9063_v109_revW_14Aug2019.

Naravaneni, R., and Jamil, K. (2005). Rapid detection of food-borne pathogens by using molecular techniques. J. Med. Microbiol. 54, 51–54. doi:10.1099/jmm.0.45687-0.

Nasheri, N., Petronella, N., Ronholm, J., Bidawid, S., and Corneau, N. (2017). Characterization of the genomic diversity of norovirus in linked patients using a metagenomic deep sequencing approach. Front. Microbiol. 8, 1–14. doi:10.3389/fmicb.2017.00073.

Negida, A., Fahim, N. K., and Negida, Y. (2019). Sample Size Calculation Guide - Part 4: How to Calculate the Sample Size for a Diagnostic Test Accuracy Study based on Sensitivity, Specificity, and the Area Under the ROC Curve. Adv. J. Emerg. Med. 3. doi:10.22114/ajem.v0i0.158.

Nieminen, T. T., Koskinen, K., Laine, P., Hultman, J., Säde, E., Paulin, L., et al. (2012). Comparison of microbial communities in marinated and unmarinated broiler meat by metagenomics. Int. J. Food Microbiol. 157, 142–149. doi:10.1016/j.ijfoodmicro.2012.04.016.

Nishikawa, Y., Kogawa, M., Hosokawa, M., Wagatsuma, R., Mineta, K., Takahashi, K., et al. (2022). Validation of the application of gel beads-based single-cell genome sequencing platform to soil and seawater. ISME Commun. 2, 92. doi:10.1038/s43705-022-00179-4.

Nouws, S., Bogaerts, B., Verhaegen, B., Denayer, S., Crombé, F., De Rauw, K., et al. (2020a). The Benefits of Whole Genome Sequencing for Foodborne Outbreak Investigation from the Perspective of a National Reference Laboratory in a Smaller Country. Foods 9, 1030. doi:10.3390/foods9081030.

Nouws, S., Bogaerts, B., Verhaegen, B., Denayer, S., Laeremans, L., Marchal, K., et al. (2021). Whole Genome Sequencing Provides an Added Value to the Investigation of Staphylococcal Food Poisoning Outbreaks. Front. Microbiol. 12, 1–16. doi:10.3389/fmicb.2021.750278.

Nouws, S., Bogaerts, B., Verhaegen, B., Denayer, S., Piérard, D., Marchal, K., et al. (2020b). Impact of DNA extraction on whole genome sequencing analysis for characterization and relatedness of Shiga toxin-producing *Escherichia coli* isolates. Sci. Rep. 10, 14649. doi:10.1038/s41598-020-71207-3.

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

Okabe, A., and Kaneda, A. (2023). Hi-C Analysis to Identify Genome-Wide Chromatin Structural Aberration in Cancer. Methods Mol Biol 2519, 127–140. doi:10.1007/978-1-0716-2433-3_15.

Okoh, A. I., Sibanda, T., and Gusha, S. S. (2010). Inadequately Treated Wastewater as a Source of Human Enteric Viruses in the Environment. Int. J. Environ. Res. Public Health 7, 2620–2637. doi:10.3390/ijerph7062620.

Okonechnikov, K., Conesa, A., and García-Alcalde, F. (2015). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. doi:10.1093/bioinformatics/btv566.

Oldach, D. W., Richard, R. E., Borza, E. N., and Benitez, R. M. (1998). A Mysterious Death. N. Engl. J. Med. 338, 1764–1769. doi:10.1056/NEJM199806113382411.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., et al. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 17, 132. doi:10.1186/s13059-016-0997-x.

Oxford Nanopore technologies (2019). Ligation sequencing gDNA - whole genome amplification protocol.

P Deloukas, Matthews, L. H., Ashurst, J., Burton, J., Gilbert, J. G., Jones, M., et al. (2001). The DNA sequence and comparative analysis of human chromosome 20. Nature 414, 865–871. doi:10.1038/414865a.

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

Park, E. J., Kim, K. H., Abell, G. C. J., Kim, M. S., Roh, S. W., and Bae, J. W. (2011). Metagenomic analysis of the viral communities in fermented foods. Appl. Environ. Microbiol. 77, 1284–1291. doi:10.1128/AEM.01859-10.

Pasteur, L., Joubert, J., and Chamberland, C. (1879). On the Germ Theory. Edinb. Med. J. 25, 265–268. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29640348.

Patravale, V., Dandekar, P., and Jain, R. (2012). Nanoparticulate Drug Delivery: Perspectives On The Transition From Laboratory To Market. Oxford: Woodhead Pub.

Payne, A., Holmes, N., Rakyan, V., and Loose, M. (2018). Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. bioRxiv. doi:https://doi.org/10.1101/312256.

Peterson, C.-L., Alexander, D., Chen, J. C.-Y., Adam, H., Walker, M., Ali, J., et al. (2022). Clinical Metagenomics Is Increasingly Accurate and Affordable to Detect Enteric Bacterial Pathogens in Stool. Microorganisms 10, 441. doi:10.3390/microorganisms10020441.

Petronella, N., Ronholm, J., Suresh, M., Harlow, J., Mykytczuk, O., Corneau, N., et al. (2018). Genetic characterization of norovirus GII.4 variants circulating in Canada using a metagenomic technique. BMC Infect. Dis. 18, 1–11. doi:10.1186/s12879-018-3419-8.

Pijnacker, R., Dallman, T. J., Tijsma, A. S. L., Hawkins, G., Larkin, L., Kotila, S. M., et al. (2019). An international outbreak of *Salmonella enterica* serotype Enteritidis linked to eggs from Poland: a microbiological and epidemiological study. Lancet Infect. Dis. 19, 778–786. doi:10.1016/S1473-3099(19)30047-7.

Pradhan, P., and Tamang, J. P. (2019). Phenotypic and Genotypic Identification of Bacteria Isolated From Traditionally Prepared Dry Starters of the Eastern Himalayas. Front. Microbiol. 10. doi:10.3389/fmicb.2019.02526.

Prestinaci, F., Pezzotti, P., and Pantosti, A. (2015). Antimicrobial resistance: A global multifaceted phenomenon. Pathog. Glob. Health 109, 309–318. doi:10.1179/2047773215Y.0000000030.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490, 55–60. doi:10.1038/nature11450.

Qiu, Q., Wang, J., Yan, Y., Roy, B., Chen, Y., Shang, X., et al. (2020). Metagenomic Analysis Reveals the Distribution of Antibiotic Resistance Genes in a Large-Scale Population of Healthy Individuals and Patients With Varied Diseases. Front. Mol. Biosci. 7. doi:10.3389/fmolb.2020.590018.

Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., et al. (2015). Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. Genome Biol. 16, 1–14. doi:10.1186/s13059-015-0677-2.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. Nat. Biotechnol. 35, 833–844. doi:10.1038/nbt.3935.

Rajagopala, S. V., Bakhoum, N. G., Pakala, S. B., Shilts, M. H., Rosas-Salazar, C., Mai, A., et al. (2021). Metatranscriptomics to characterize respiratory virome, microbiome, and host response directly from clinical samples. Cell Reports Methods 1, 100091. doi:10.1016/j.crmeth.2021.100091.

Revez, J., Espinosa, L., Albiger, B., Leitmeyer, K. C., and Struelens, M. J. (2017). Survey on the Use of Whole-Genome Sequencing for Infectious Diseases Surveillance: Rapid Expansion of European National Capacities, 2015–2016. Front. Public Heal. 5. doi:10.3389/fpubh.2017.00347.

Robert Koch Institute (2011). Report: Final presentation and evaluation of epidemiological findings in the EHEC O104 : H4 outbreak, Germany 2011.

Rose, G., Wooldridge, D. J., Anscombe, C., Mee, E. T., Misra, R. V., and Gharbia, S. (2015). Challenges of the Unknown: Clinical Application of Microbial Metagenomics. Int. J. Genomics 2015. doi:10.1155/2015/292950.

Russell, J. A., Campos, B., Stone, J., Blosser, E. M., Burkett-Cadena, N., and Jacobs, J. L. (2018). Unbiased Strain-Typing of Arbovirus Directly from Mosquitoes Using Nanopore Sequencing: A Field-forward Biosurveillance Protocol. Sci. Rep. 8, 1–12. doi:10.1038/s41598-018-23641-7.

Sala, C., Mordhorst, H., Grützke, J., Brinkmann, A., Petersen, T. N., Poulsen, C., et al. (2020). Metagenomics-based proficiency test of smoked salmon spiked with a mock community. Microorganisms 8, 1–16. doi:10.3390/microorganisms8121861.

Saltykova, A. (2022). Development, evaluation and optimization of next generation sequencing data analysis tools in support of a fast response for public health and food chain safety. Ghent University. Faculty of Sciences, Ghent, Belgium.

Saltykova, A., Buytaers, F. E., Denayer, S., Verhaegen, B., Piérard, D., Roosens, N. H. C., et al. (2020). Strain-Level Metagenomic Data Analysis of Enriched In Vitro and In Silico Spiked Food Samples: Paving the Way towards a Culture-Free Foodborne Outbreak Investigation Using STEC as a Case Study. Int. J. Mol. Sci. 21, 5688. doi:10.3390/ijms21165688.

Saltykova, A., Wuyts, V., Mattheus, W., Bertrand, S., Roosens, N. H. C., Marchal, K., et al. (2018). Comparison of SNP-based subtyping workflows for bacterial isolates using WGS data, applied to *Salmonella enterica* serotype Typhimurium and serotype 1,4,[5],12:i:-. PLoS One 13, 1–23. doi:10.1371/journal.pone.0192504.

Sandora, T. J., Gerner-Smidt, P., and McAdam, A. J. (2014). What's your subtype ? The epidemiologic utility of bacterial whole-genome sequencing. Clin. Chem. 60, 586–588. doi:10.1373/clinchem.2013.217141.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. 74, 5463–5467. doi:10.1073/pnas.74.12.5463.

Santos, I. C., Smuts, J., Choi, W.-S., Kim, Y., Kim, S. B., and Schug, K. A. (2018). Analysis of bacterial FAMEs using gas chromatography – vacuum ultraviolet spectroscopy for the identification and discrimination of bacteria. Talanta 182, 536–543. doi:10.1016/j.talanta.2018.01.074.

Sarkar, A., Harty, S., Moeller, A. H., Klein, S. L., Erdman, S. E., Friston, K. J., et al. (2021). The gut microbiome as a biomarker of differential susceptibility to SARS-CoV-2. Trends Mol. Med. 27, 1115–1134. doi:10.1016/j.molmed.2021.09.009.

Sayers, E. W., Agarwala, R., Bolton, E. E., Brister, J. R., Canese, K., Clark, K., et al. (2019). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 47, D23–D28. doi:10.1093/nar/gky1069.

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., et al. (2022). Database resources of the national center for biotechnology information. Nucleic Acids Res. 50, D20–D26. doi:10.1093/nar/gkab1112.

Scallan, E., Hoekstra, R. M., Angulo, F. J., Tauxe, R. V., Widdowson, M. A., Roy, S. L., et al. (2011). Foodborne illness acquired in the United States-Major pathogens. Emerg. Infect. Dis. 17, 7–15. doi:10.3201/eid1701.P11101.

Scheuch, M., Höper, D., and Beer, M. (2015). RIEMS: A software pipeline for sensitive and comprehensive taxonomic classification of reads from metagenomics datasets. BMC Bioinformatics 16. doi:10.1186/s12859-015-0503-6.

Schmidt, K., Mwaigwisya, S., Crossman, L. C., Doumith, M., Munroe, D., Pires, C., et al. (2017). Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. J. Antimicrob. Chemother. 72, 104–114. doi:10.1093/jac/dkw397.

Schöpflin, R., Melo, U. S., Moeinzadeh, H., Heller, D., Laupert, V., Hertzberg, J., et al. (2022). Integration of Hi-C with short and long-read genome sequencing reveals the structure of germline rearranged genomes. Nat. Commun. 13, 6470. doi:10.1038/s41467-022-34053-7.

Sciensano (2020). Voedselvergiftigingen in Belgie jaaroverzicht 2019. Available at: https://www.sciensano.be/sites/default/files/jaarverslagboekje_vti2019_nl2020.pdf.

Seeman, T. (2015). snippy: fast bacterial variant calling from NGS reads. Available at: https://github.com/tseemann/snippy.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069. doi:10.1093/bioinformatics/btu153.

Sekse, C., Holst-Jensen, A., Dobrindt, U., Johannessen, G. S., Li, W., Spilsberg, B., et al. (2017). High throughput sequencing for detection of foodborne pathogens. Front. Microbiol. 8, 1–26. doi:10.3389/fmicb.2017.02029.

Selander, R. K., Caugant, D. A., Ochman, H., Musser, J. M., Gilmour, M. N., and Whittam, T. S. (1986). Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. Appl. Environ. Microbiol. 51, 873–884. doi:10.1128/aem.51.5.873-884.1986.

Shah, S. J., Barish, P. N., Prasad, P. A., Kistler, A., Neff, N., Kamm, J., et al. (2020). Clinical features, diagnostics, and outcomes of patients presenting with acute respiratory illness: A retrospective cohort study of patients with and without COVID-19. EClinicalMedicine 27, 100518. doi:10.1016/j.eclinm.2020.100518.

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., et al. (2017). DNA sequencing at 40: Past, present and future. Nature 550. doi:10.1038/nature24286.

Silano, V., Barat Baviera, J. M., Bolognesi, C., Brüschweiler, B. J., Cocconcelli, P. S., Crebelli, R., et al. (2019). Characterisation of microorganisms used for the production of food enzymes. EFSA J. 17, 1–13. doi:10.2903/j.efsa.2019.5741.

Sinclair, J. R. (2019). Importance of a One Health approach in advancing global health security and the Sustainable Development Goals. Rev. Sci. Tech. l'OIE 38, 145–154. doi:10.20506/rst.38.1.2949.

Singh, N., Lapierre, P., Quinlan, T. M., Halse, T. A., Wirth, S., Dickinson, M. C., et al. (2019). Whole_Genome Single-Nucleotide Polymorphism (SNP) Analysis Applied Directly to Stool for Genotyping Shiga Toxin-Producing *Escherichia coli*: an Advanced Molecular detection method for foodborne disease surveillance and outbreak tracking. J. Clin. Microbiol. 57, 1–11.

Snow, J. (1856). Cholera and the Water Supply in the South Districts of London in 1854. J. public Heal. Sanit. Rev. 2, 239–257. Available at: http://www.ncbi.nlm.nih.gov/pubmed/30378891.

Somerville, V., Lutz, S., Schmid, M., Frei, D., Moser, A., Irmler, S., et al. (2018). Long read-based de novo assembly of low complex metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. bioRxiv, 476747. doi:10.1101/476747.

Somura, Y., Nagano, M., Kimoto, K., Oda, M., Mori, K., Shinkai, T., et al. (2019). Detection of norovirus in food samples collected during suspected food-handler-involved foodborne outbreaks in Tokyo. Lett. Appl. Microbiol., lam.13189. doi:10.1111/lam.13189.

Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H., and DeLong, E. F. (1996). Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. J. Bacteriol. 178, 591–599. doi:10.1128/jb.178.3.591-599.1996.

Stevens, E. J., Bates, K. A., and King, K. C. (2021). Host microbiota can facilitate pathogen infection. PLOS Pathog. 17, e1009514. doi:10.1371/journal.ppat.1009514.

Strain, R., Stanton, C., and Ross, R. P. (2022). Effect of diet on pathogen performance in the microbiome. Microbiome Res. Reports. doi:10.20517/mrr.2021.10.

Strubbia, S., Phan, M. V. T., Schaeffer, J., Koopmans, M., Cotten, M., and Le Guyader, F. S. (2019). Characterization of Norovirus and Other Human Enteric Viruses in Sewage and Stool Samples Through Next-Generation Sequencing. Food Environ. Virol. 11, 400–409. doi:10.1007/s12560-019-09402-3.

Strubbia, S., Schaeffer, J., Besnard, A., Wacrenier, C., Le Mennec, C., Garry, P., et al. (2020). Metagenomic to evaluate norovirus genomic diversity in oysters: Impact on hexamer selection and targeted capture-based enrichment. Int. J. Food Microbiol. 323, 108588. doi:10.1016/j.ijfoodmicro.2020.108588.

Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10. doi:10.1093/oxfordjournals.molbev.a040023.

Tang, B., and Row, K. H. (2013). Development of Gas Chromatography Analysis of Fatty Acids in Marine Organisms. J. Chromatogr. Sci. 51, 599–607. doi:10.1093/chromsci/bmt005.

Tang, S., Orsi, R. H., Luo, H., Ge, C., Zhang, G., Baker, R. C., et al. (2019). Assessment and comparison of molecular subtyping and characterization methods for *Salmonella*. Front. Microbiol. 10. doi:10.3389/fmicb.2019.01591.

Tang, Y. W., Ellis, N. M., Hopkins, M. K., Smith, D. H., Dodge, D. E., and Persing, D. H. (1998). Comparison of phenotypic and genotypic techniques for identification of unusual aerobic pathogenic gram-negative *bacilli*. J. Clin. Microbiol. 36, 3674–3679. doi:10.1128/jcm.36.12.3674-3679.1998.

Thung, T. Y., Radu, S., Mahyudin, N. A., Rukayadi, Y., Zakaria, Z., Mazlan, N., et al. (2018). Prevalence, virulence genes and antimicrobial resistance profiles of *Salmonella* serovars from retail beef in Selangor, Malaysia. Front. Microbiol. 8, 1–8. doi:10.3389/fmicb.2017.02697.

Timme, R. E., Rand, H., Leon, M. S., Hoffmann, M., Strain, E., Allard, M., et al. (2018). GenomeTrakr proficiency testing for foodborne pathogen surveillance: An exercise from 2015. Microb. Genomics 4. doi:10.1099/mgen.0.000185.

Tomayko, J. F., and Murray, B. E. (1995). Analysis of *Enterococcus faecalis* isolates from intercontinental sources by multilocus enzyme electrophoresis and pulsed-field gel electrophoresis. J. Clin. Microbiol. 33, 2903–2907. doi:10.1128/jcm.33.11.2903-2907.1995.

Towner, K. J., and Cockayne, A. (1993). “Typing and identification of microorganisms with antibodies,” in Molecular Methods for Microbial Identification and Typing (Dordrecht: Springer Netherlands), 159–186. doi:10.1007/978-94-011-1506-3_6.

Tran, Q., and Phan, V. (2020). Assembling Reads Improves Taxonomic Classification of Species. Genes 11, 946. doi:10.3390/genes11080946.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., et al. (2009). A core gut microbiome in obese and lean twins. Nature 457, 480–484. doi:10.1038/nature07540.

U.S. Department of Agriculture, A. R. S. (2020). USDA Food and Nutrient Database for Dietary Studies 2017-2018. Food Surveys Research Group Home Page, http://www. Available at: http://www.ars.usda.gov/nea/bhnrc/fsrg.

UE (2005). COMMISSION REGULATION (EC) No 2073/2005 of 15 November 2005 on microbiological criteria for foodstuffs. Off. J. Eur. Union 32.

Uelze, L., Becker, N., Borowiak, M., Busch, U., Dangel, A., Deneke, C., et al. (2021). Toward an Integrated Genome-Based Surveillance of *Salmonella enterica* in Germany. Front. Microbiol. 12. doi:10.3389/fmicb.2021.626941.

Uyttendaele, M. (2020). “UGent Course Food Safety & Risk Analysis, Partim – Microbial hazards,” in Microbiological guidelines.

Van Beek, J., De Graaf, M., Smits, S., Schapendonk, C. M. E., Verjans, G. M. G. M., Vennema, H., et al. (2017). Whole-Genome Next-Generation Sequencing to Study Within-Host Evolution of Norovirus (NoV) among Immunocompromised Patients with Chronic NoV Infection. J. Infect. Dis. 216, 1513–1524. doi:10.1093/infdis/jix520.

Van Goethem, N., Struelens, M. J., De Keersmaecker, S. C. J., Roosens, N. H. C., Robert, A., Quoilin, S., et al. (2020). Perceived utility and feasibility of pathogen genomics for public health practice: a survey among public health professionals working in the field of infectious diseases, Belgium, 2019. BMC Public Health 20, 1318. doi:10.1186/s12889-020-09428-4.

Veloo, A. C. M., Jean-Pierre, H., Justesen, U. S., Morris, T., Urban, E., Wybo, I., et al. (2018). Validation of MALDI-TOF MS Biotyper database optimized for anaerobic bacteria: The ENRIA project. Anaerobe 54, 224–230. doi:10.1016/j.anaerobe.2018.03.007.

Verhaegen, B., De Reu, K., Heyndrickx, M., and De Zutter, L. (2015). Comparison of Six Chromogenic Agar Media for the Isolation of a Broad Variety of Non-O157 Shigatoxin-Producing *Escherichia coli* (STEC) Serogroups. Int. J. Environ. Res. Public Health 12, 6965–6978. doi:10.3390/ijerph120606965.

Veziant, J., Villéger, R., Barnich, N., and Bonnet, M. (2021). Gut Microbiota as Potential Biomarker and/or Therapeutic Target to Improve the Management of Cancer: Focus on Colibactin-Producing *Escherichia coli* in Colorectal Cancer. Cancers (Basel). 13, 2215. doi:10.3390/cancers13092215.

Vibin, J., Chamings, A., Collier, F., Klaassen, M., Nelson, T. M., and Alexandersen, S. (2018). Metagenomics detection and characterisation of viruses in faecal samples from Australian wild birds. Sci. Rep. 8, 1–23. doi:10.1038/s41598-018-26851-1.

Vivancos, R., Shroufi, A., Sillis, M., Aird, H., Gallimore, C. I., Myers, L., et al. (2009). Food-related norovirus outbreak among people attending two barbeques: epidemiological, virological, and environmental investigation. Int. J. Infect. Dis. 13, 629–635. doi:10.1016/j.ijid.2008.09.023.

Volk, H., Piskernik, S., Kurincic, M., Klancnik, A., Toplak, N., and Jersek, B. (2014). Evaluation of different methods for DNA extraction from milk. J. Food Nutr. Res. 53, 97–104. Available at: http://journal.hibiscuspublisher.com/index.php/JOBIMB/article/view/150.

Walsh, A. M., Crispie, F., Daari, K., O'Sullivan, O., Martin, J. C., Arthur, C. T., et al. (2017). Strain-Level Metagenomic Analysis of the Fermented Dairy Beverage Nunu Highlights Potential Food Safety Risks. Appl. Environ. Microbiol. 83, 1–13. doi:10.1128/AEM.01144-17.

Wang, W.-L., Xu, S.-Y., Ren, Z.-G., Tao, L., Jiang, J.-W., and Zheng, S.-S. (2015). Application of metagenomics in the human gut microbiome. World J. Gastroenterol. 21, 803. doi:10.3748/wjg.v21.i3.803.

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., and Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. Nat. Biotechnol. 39, 1348–1365. doi:10.1038/s41587-021-01108-x.

Werber, D., Krause, G., Frank, C., Fruth, A., Flieger, A., Mielke, M., et al. (2012). Outbreaks of virulent diarrheagenic *Escherichia coli* - are we in control? BMC Med. 10, 11. doi:10.1186/1741-7015-10-11.

Wetterstrand KA DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at: www.genome.gov/sequencingcostsdata [Accessed June 22, 2022].

WHO (2015). WHO Estimates of the Global Burden of foodborne diseases. Available at: https://apps.who.int/iris/bitstream/handle/10665/199350/9789241565165_eng.pdf?sequence=1.

WHO (2018). Whole genome sequencing for foodborne disease surveillance - landscape paper. ISBN:978-92-4-151386-9

Williams, rachel J., Tutill, H., Roy, S., Romero, E. Y., Williams, C. A., and Breuer, J. (2019). Agilent Application Note: Utilization of Agilent SureSelect Target Enrichment for Whole Genome Sequencing of Viruses and Bacteria.

Winand, R., Bogaerts, B., Hoffman, S., Lefevre, L., Delvoye, M., Van Braekel, J., et al. (2019). Targeting the 16S rRNA Gene for Bacterial Identification in Complex Mixed Samples: Comparative Evaluation of Second (Illumina) and Third (Oxford Nanopore Technologies) Generation Sequencing Technologies. Int. J. Mol. Sci. 21, 298. doi:10.3390/ijms21010298.

Woese, C. R., and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc. Natl. Acad. Sci. 74, 5088–5090. doi:10.1073/pnas.74.11.5088.

Wongsurawat, T., Jenjaroenpun, P., Taylor, M. K., Lee, J., Tolardo, A. L., Parvathareddy, J., et al. (2019). Rapid sequencing of multiple RNA viruses in their native form. Front. Microbiol. 10, 1–8. doi:10.3389/fmicb.2019.00260.

Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. bioRxiv, 762302. doi:10.1101/762302.

Wood, D. E., and Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15. doi:10.1186/gb-2014-15-3-r46.

Wright, R. J., Comeau, A. M., and Langille, morgan G. I. (2022). From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools. bioRxiv. doi:10.1101/2022.04.27.489753.

Xu, Y., and Zhao, F. (2018). Single-cell metagenomics: challenges and applications. Protein Cell 9, 501–510. doi:10.1007/s13238-018-0544-5.

Yang, X., Noyes, N. R., Doster, E., Martin, J. N., Linke, L. M., Magnuson, R. J., et al. (2016). Use of metagenomic shotgun sequencing technology to detect foodborne pathogens within the microbiome of the beef production chain. Appl. Environ. Microbiol. 82, 2433–2443. doi:10.1128/AEM.00078-16.

Yang, Z., Mammel, M., Papafragkou, E., Hida, K., Elkins, C. A., and Kulka, M. (2017). Application of next generation sequencing toward sensitive detection of enteric viruses isolated from celery samples as an example of produce. Int. J. Food Microbiol. 261, 73–81. doi:10.1016/j.ijfoodmicro.2017.07.021.

Ye, S. H., Siddle, K. J., Park, D. J., and Sabeti, P. C. (2019). Benchmarking Metagenomics Tools for Taxonomic Classification. Cell 178, 779–794. doi:10.1016/j.cell.2019.07.010.

Yoshida, C. E., Kruczkiewicz, P., Laing, C. R., Lingohr, E. J., Gannon, V. P. J., Nash, J. H. E., et al. (2016). The *salmonella* in silico typing resource (SISTR): An open web-accessible tool for rapidly typing and subtyping draft *salmonella* genome assemblies. PLoS One 11, 1–17. doi:10.1371/journal.pone.0147101.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., et al. (2012). Identification of acquired antimicrobial resistance genes. J. Antimicrob. Chemother. 67, 2640–2644. doi:10.1093/jac/dks261.

Zhang, W., Li, L., Deng, X., Kapusinszky, B., and Delwart, E. (2014). What is for dinner? Viral metagenomics of US store bought beef, pork, and chicken. Virology 468–470, 303–310. doi:10.1016/j.virol.2014.08.025.

Zhou, X., Ren, L., Meng, Q., Li, Y., Yu, Y., and Yu, J. (2010). The next-generation sequencing technology and application. Protein Cell 1, 520–536. doi:10.1007/s13238-010-0065-3.

Zhou, Z., Luhmann, N., Alikhan, N.-F., Quince, C., and Achtman, M. (2018). "Accurate Reconstruction of Microbial Strains from Metagenomic Sequencing Using Representative Reference Genomes," in, 225–240. doi:10.1007/978-3-319-89929-9_15.

Curriculum vitae

Personal information

Name	Buytaers Florence
Date of birth	08/10/1992
Place of birth	Brussels, Belgium

Education

2018 – now	**Doctor of Science: Biochemistry and Biotechnology** Ghent University, Ghent, Belgium Sciensano, Brussels, Belgium
2013 – 2015	**Master in bioengineering** *Major: chemistry and bio-industries* *Minor: molecular and cellular biotechnology* ULB, Brussels, Belgium
2010– 2013	**Bachelor in engineering, direction bioengineering** ULB, Brussels, Belgium

Abstracts of posters presented at national and international conferences and workshops

- Twenty-third Conference on Food Microbiology, Brussels, October 4-5, 2018

Poster presentation: F. Buytaers, S. Denayer, B. Verhaegen, A. Saltykova, N. Roosens, K. Vanneste, D. Piérard, K. De Rauw, K. Marchal, S. C. J. De Keersmaecker. **Development of a shotgun metagenomics approach for outbreak investigation with STEC as a case study.**

- Joint conference on foodborne pathogens & whole genome sequencing, Paris, March 26-28, 2019

Poster presentation: F. Buytaers, A. Saltykova, S. Denayer, B. Verhaegen, N. Roosens, K. Vanneste, D. Piérard, K. Marchal, S. C. J. De Keersmaecker. **Comparing metagenomics, WGS on isolates and routine classical microbiology methods for foodborne outbreak investigations with STEC as a case study.**

- Nanopore Community Meeting 2020, online, December 1-3 2020

Poster presentation: F. Buytaers, A. Saltykova, M.-A. Fraiture, B. Berbers, S. Denayer, B. Verhaegen, N. Papazova, K. Vanneste, D. Piérard, N. Roosens, K. Marchal, S. C. J. De Keersmaecker. **Using a shotgun metagenomics approach with long-reads sequencing for food safety aspects.**

- ASM conference on rapid applied microbial next-generation sequencing and bioinformatics pipelines, online, December 7-11 2020

Poster presentation: F. E. Buytaers, A. Saltykova, B. Verhaegen, N. H. C. Roosens, K. Vanneste, K. Marchal, S. Denayer, S. C. J. De Keersmaecker. **Strain-Level Shotgun Metagenomics for Faster Food-Borne Outbreak Investigation.**

- Twenty-fifth Conference on Food Microbiology, Brussels, October 7-8, 2021

Poster presentation: F. E. Buytaers, A. Saltykova, B. Verhaegen, D. Piérard, W. Mattheus, N. H. C. Roosens, K. Vanneste, K. Marchal, S. Denayer, S. C. J. De Keersmaecker. **Application of Strain-Level Shotgun Metagenomics for Food-Borne Outbreak Investigation.**

- EFSA ONE conference, Brussels, June 21-24, 2022

Poster presentation: F.E. Buytaers, M. Fraiture, B. Berbers, E. Vandermassen, S. Hoffman, N. Papazova, K. Vanneste, K. Marchal, N.H.C. Roosens, S.C.J. De Keersmaecker. **A shotgun metagenomics approach to detect and characterize unauthorized genetically modified microorganisms in microbial fermentation products.**

Selected oral presentations of the work of and by the PhD student:

- Scientific Lunch, Sciensano, Brussels, October 16th, 2018: StEQIDEMIC.be.

- Scientific Lunch, Sciensano, Brussels, September 13th, 2019: Avoiding the culture step in outbreak investigations: parameters for optimised metagenomics analysis of contaminated food.

- Imeko Foods, Brussels, September 17th, 2019: Avoiding the culture step in outbreak investigations: parameters for optimised metagenomics analysis of contaminated food.

- 15th Annual workshop of the national Reference Laboratories for *E.coli* in the EU, EURL STEC, online, September 21-22, 2020: NGS for STEC characterization: the complete package for an NRL. Surveillance and outbreak investigation of isolates using a validated wet-lab and bioinformatics WGS workflow & strain-level metagenomics-based foodborne outbreak investigation and source tracking (joint presentation F. Buytaers, S. Nouws, B. Bogaerts)

- Twenty-fifth Conference on Food Microbiology, Brussels, October 7-8, 2021: Application of Strain-Level Shotgun Metagenomics for Food-Borne Outbreak Investigation.

- Midterm assessment meeting, October 20 2021: Development of strain-level shotgun metagenomics approaches to detect and characterize microbiological contaminants in the context of food safety.

- One Health EJP, online, October 27th, 2021: Strain-level shotgun metagenomics for the detection and characterization of bacterial foodborne contaminants in the context of outbreak investigation and EU regulatory enforcement (joint presentation S. De Keersmaecker)

- FoodMicro, Athens, August 28-31, 2022: Shotgun metagenomics for strain-level food-borne outbreak investigation

Publications

1. **Buytaers FE**†, Saltykova A†, Denayer S, Verhaegen B, Vanneste K, Roosens NHC, Piérard D, Marchal K, De Keersmaecker SCJ. A Practical Method to Implement Strain-Level Metagenomics-Based Foodborne Outbreak Investigation and Source Tracking in Routine, Microorganisms, 2020, 8, 1191, doi:10.3390/microorganisms8081191.

2. Saltykova A†, **Buytaers FE**†, Denayer S, Verhaegen B, Piérard D, Roosens NHC, Marchal K, De Keersmaecker SCJ. Strain-Level Metagneomic Data Analysis of Enriched In Vitro and In Silico Spiked Food Samples: paving the Way towards a Culture-Free Foodborne Outbreak investigation using STEC as a Case Study, Int. J. Mol. Sci. , 2020, 21, 5688, doi:10.3390/ijms21165688.

3. **Buytaers FE**, Saltykova A, Mattheus W, Verhaegen B, Roosens NHC, Vanneste K, Laisnez V, Hammami N, Pochet B, Cantaert V, Marchal K, Denayer S, De Keersmaecker SCJ., Application of a strain- level shotgun metagenomics approach on food samples: resolution of the source of a *Salmonella* food- borne outbreak, Microbial genomics, 2021, 7, doi:10.1099/mgen.0.000547.

4. **Buytaers FE**, Fraiture MA, Berbers B, Vandermassen E, Hoffman S, Papazova N, Vanneste K, Marchal K, Roosens NHC and De Keersmaecker SCJ. A shotgun metagenomics approach to detect and characterize unauthorized genetically modified microorganisms in microbial fermentation products, Food Chemistry: Molecular Sciences, 2021, 2, 100023, doi:10.1016/j.fochms.2021.100023.

5. **Buytaers FE**[†], Saltykova A[†], Denayer S, Verhaegen B, Vanneste K, Roosens NHC, Piérard D, Marchal K, De Keersmaecker SCJ. Towards real-time and affordable strain-level metagenomics-based food-borne outbreak investigations using Oxford Nanopore sequencing technologies, Frontiers in microbiology, 2021, 12, 1–13, doi:10.3389/fmicb.2021.738284.

6. **Buytaers FE**, Verhaegen B, Gand M, D’aes J, Vanneste K, Roosens NHC, Piérard D, Marchal K, Denayer S, De Keersmaecker SCJ. Metagenomics to Detect and Characterize Viruses in Food Samples at Genome Level? Lessons Learnt from a Norovirus Study. Foods 2022, 11, 3348, doi:10.3390/foods11213348.

[†] Contributed equally

Acknowledgements

Since I was a little girl, I have been drawn to science with the desire to understand the world around me. My parents, as well as my grandparents, have always believed in me and pushed me to seek answers to my questions, with a critical mind. They always supported me throughout my studies and career even when I decided to leave to the other side of the world. My brother and friends also accompanied me happily through all these years by always being one phone call away and knowing how to make me laugh when my mind needed a rest.

When I met Remy, I met my husband, but also someone who is always there for me and who also believed in my desire to do scientific research. In 2018, when I was wondering which position would fit me better, he pushed me to embark on the completion of a PhD at Sciensano, the profession that has fulfilled me for the last four years. During this time, always looking for adventure in our life, we have travelled, bought and renovated a flat, and finally welcomed our little Daphne, who with her bright smiles has highlighted our days and made me want to set the best example as a scientist mum. Through meeting Remy, I also met a second family, with members from Belgium and much further away, who welcomed me with open arms and always admired my desire to do a PhD and work to achieve my dreams.

Through this PhD, I also discovered a new family in my work environment. First of all, Sigrid, my wonderful promoter who accompanied me day after day (and sometimes evenings as well) at Sciensano. She always had a listening ear during our meetings and could bring fresh ideas and energy to keep the project going when nothing would work as expected. She always took time to accompany me through all the stages of each of my papers as well as this thesis. I would also like to thank the rest of the TAG team, especially Nancy, the head of our service, member of my accompanying committee, and co-author of my publications, who always listened to my progress and involved me in research on GMMs. Then of course come the other PhD students and postdocs, with whom we formed a close team from the beginning, the laboratory technicians, who also made this work possible, and the bioinformaticians, for their wise advices. But at Sciensano I did not only work at TAG, I also worked closely with the food pathogens, where Sarah was always present to share her expertise, protocols, laboratory and samples. She accompanied me throughout my PhD as a member of the accompanying committee and co-author of most of my papers. I would also like to thank Bavo and the laboratory technicians from the food pathogens, who always took time to answer my questions or get something for me. I also had the chance to work with Wesley from the bacterial diseases, and Denis from the UZ VUB for the human samples in my studies. Kathleen, my promoter from UGent, made this all possible and was always present to share her expertise and encouragements.

Finally, I also want to thank the members of the jury for the time they took to read this thesis, be present at my defenses and share insightful comments.

Thank you all, this is not the work of just one person, but an accomplishment made possible with a team!

FACULTY OF SCIENCES

DEPARTMENT OF PLANT BIOTECHNOLOGY AND BIOINFORMATICS

Krijgslaan 281 - S2
9000 GHENT, BELGIUM
https://www.ugent.be/we