Applications of Next Generation Sequencing for the Assessment of Microbiological Safety of Fresh Produce

Erin Rachel Lewis

Doctor of Philosophy

School of Natural and Environmental Science,

Newcastle University

November 2019

Abstract

Fresh produce, which is often minimally processed and consumed raw, is increasingly recognised as a route of contamination leading to foodborne illness. This study focussed on the potential of next generation sequencing (NGS) to detect and characterise foodborne pathogens and elucidate the microbiome and potential influences on the survival and transmission of human pathogens within the fresh produce supply-chain.

Initial research assessed the ability of several NGS methodologies to screen the microbiome of fresh produce and identify the limit of detection of bacterial and viral pathogens, using a mock sample set created with known levels of contamination. The limit of detection of human pathogens was found to be dependent upon enrichment method, sequencing approach and bioinformatics analysis utilised. The most sensitive approach tested involved sequence preparation using ribosomal depletion followed by RNAseq and analysis of the microbiome using Kraken; yielding a limit of detection of 10^5 colony forming units / plaque forming units (CFU/PFU) per extraction. Subsequent *in silico* work showed the differing read lengths obtained from the MiSeq, HiSeq and NovaSeq has no influence on the limit of detection of human pathogens within a mock community.

Methodological work was applied to study the microbiome associated with commercial fresh produce samples to identify species that may confer a positive or negative effect on the survival of human pathogens. Sequences were also screened for antimicrobial resistance (AMR) associated genes. Data revealed the microbiome to be dominated by potential spoilage organisms and plant pathogens. Four unique AMR-associated genes, with over 80% identity and coverage, were found (*CRP, H-NS, MexF, MexB*). Nine taxa, including *Pectobacterium* and *Dickeya* - soft rot causing bacteria previously linked in the literature to survival of *Salmonella* on fresh produce - were found to be positively correlated with the microbiological detection of Enterobacteriaceae and *Listeria* species. *Pseudonocardia* was the only taxon detected in a large proportion of samples that was inversely correlated with detection of Enterobacteriaceae.

A collection of 48 strains of *Listeria monocytogenes* originating from the fresh produce supply chain was subject to a combination of phenotypic and genotypic methods to characterise the resistome, virulome, and biofilm forming ability. The genes present and phylogenetic identity of these isolates was then compared to those of 80 isolates from meat

and clinical origin, to identify signatures of fresh produce contaminating *L. monocytogenes*. Phenotypic screens revealed no evidence of AMR within *L. monocytogenes* isolated from UK fresh produce. All isolates of *L. monocytogenes* tested were capable of forming biofilms and displayed an increased propensity for biofilm formation in nutrient broth than in brain heart infusion broth (suggesting that biofilm formation may be induced by stress). Whole genome sequencing data from fresh produce isolates, meat isolates and clinical isolates were indistinguishable based on phylogeny, resistome and virulome. When screened using genome wide association study (GWAS) many potential genes of interest were highlighted.

Acknowledgements

To my supervisors. Jerry Barnes, for his honesty and advice. Andrew Hudson and Nigel Cook, for remaining with me throughout my project despite no longer being paid to do so. Edward Haynes, for his care and support through a difficult three years. He talked me down on more occasions than I can count and kept me going when I thought it was not possible. Thank you so much.

To Nicola Wilson, a special thank you for her help and determination in the provision of both samples and information. Without her this thesis would be a very different piece of work.

To my colleagues and friends at Fera, both past and present. There are too many people to name you all. To Melanie Sapp, you built me up and gave me the opportunity to be who I am today. To Ian Adams, Sam McGreig, and John Walshaw, for letting me test their knowledge so thoroughly and help me whenever I needed. To Lucy Vickers-Smith, Ioana Lock, Emiline Quill and Jayne Hall, for lab support and spending long hours in a small room with me!

To my friends. Clare, thank you for reminding me I was not alone. Sam, you get a double mention! Jenny and Ellie, for impartial advice and the occasional present to cheer me up. Amy, for always listening. To those of you too numerous to name who have listened and been amazing friends.

And finally, to my mum and sister. Thank you for putting up with tantrums and breakdowns. Thank you for the final typo checking at fast turnaround! You are my rocks and my world. I love you both.

Collaborations

All work presented herein is the author's own, with the following exceptions:

Chapter 2: Extraction of sample from Fera archive was undertaken by Lucy Vickers-Smith and Nigel Cook as part of a previous project.

Chapter 3: Microbiological sampling, testing and results were provided by Westward Laboratories.

Chapter 4: WGS of isolates veg 1-14 was performed by Alva Smith at Edinburgh Napier University as part of his PhD with the original isolates and raw fastqs provided for inclusion in this thesis.

Contents

Lists of Figures

List of Tables

Chapter 1. General Introduction

1.1 Foodborne Disease

The World Health Organisation (2015b) estimated that there were 600 million cases of foodborne illness in 2010 and over 1 million people per year were affected in the UK alone. This costs the UK economy nearly £1.5 bn in healthcare and associated costs (Food Standards Agency 2011). The first report that the consumption of contaminated produce may result in disease/food poisoning was in 1912, when leafy greens were recognised as a potential vector for *Bacillus typhosus* (Creel 1912). In the past century, there has been a rise in the demand for fresh produce driven by the recognition that fruit, salads and vegetables are important constituents of a healthy diet, and through marketing campaigns in recent years promoting healthy eating and a 5-a-day rhetoric (Allen et al. 2013). Moreover, consumer demands for fresh produce all-year-round have led to increases in transportation, faster throughput production and intensive farming approaches (Chitarra et al. 2014; Fatica and Schneider 2011) which have increased the risk of contamination by human pathogens (Ziuzina et al. 2015). The fact that fresh produce invariably has a short shelf-life and is commonly eaten raw and/or with minimal processing means that there is a substantially enhanced risk of low-level contamination by several potential pathogens. Moreover, for this type of short-shelf life produce routine microbiological testing approaches are inappropriate since the results are generally returned several days after the products are consumed and thus too late to instigate safe practices, such as withdrawal from market. It is predicted that ≈10% of food poisoning cases reported in the UK are attributable to contamination of fresh produce and/or minimally-processed foods (Tam *et al.* 2014). Reported incidences of food poisoning associated with “salads” are lower at just 4% of reported UK cases; impacting ≈3,500 people in 82 separate outbreaks (Little and Gillespie 2008). It seems likely that any such statistics underestimate the health issues associated with fresh and/or minimally-processed produce due to (i) the inherent short shelf life of such products which is commonly shorter than the incubation time of the disease agent (ii) uncertainties in recall during epidemiological studies - people often forget that fresh produce was ingested as a garnish or a side as part of other meals (iii) a relatively small proportion of food poisoning cases are investigated in any detail and (iv) it is becoming increasingly difficult to identify and track outbreaks of food poisoning due to the burgeoning market for the trans-national shipment of foods from large centralised processing plants (Monaghan *et al.* 2008).

1.1.1 Diversity of Pathogens causing Foodborne Disease

There are a wide range of bacterial, fungal, viral and protozoan human pathogens that can lead to illness following the consumption of fresh produce. In the UK and Europe, outbreaks of identified aetiology associated with fresh produce were predominantly caused by bacteria (Little and Gillespie 2008; European Food Safety Authority 2016). The major identified causative agents are outlined in Table 1 and described in greater detail below. Of the foodborne outbreaks (FBO) identified in the EU in 2017, 33% were of unknown causes (European Food Safety Authority 2018).

Table 1. Case, hospitalisation, death and foodborne associated cases and outbreaks for the top four foodborne associated bacteria in 2017

Pathogen	Total Disease[a]			Food-borne disease[a]	
	Human Cases	Hospitalisation	Deaths	Human Cases	Outbreaks
Campylobacter	246,158	20,810	45	1,445	395
Salmonella	91,662	16,796	156	9,600	1,241
STEC[b]	6,073	933	20	206	48
Listeria	2,480	988	225	39	10

[a] Information adapted from the European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2017 (European Food Safety Authority 2018).
[b] Shiga toxin-producing *E. coli*

1.1.1.1 Salmonella

Salmonella are rod-shaped gram-negative Enterobacteriaceae. There are two species of *Salmonella: Salmonella enterica* and *Salmonella bongori*. *Salmonella enterica* is divided into 6 sub-species. There are currently 2659 described serovars of *Salmonella* recorded (Issenhuth-Jeanjean *et al.* 2014). *Salmonella enterica* can cause two types of disease, typhoid fever and food poisoning (Boyle *et al.* 2007). Symptoms of non-typhoidal *Salmonella* infection include diarrhoea, stomach cramps, vomiting and fever. The incubation period for the disease is 12 - 72 hours and symptoms last four to seven days (NHS 2014). Most cases of infection with *Salmonella* are mild and usually do not require treatment. However, in some cases severe dehydration can result in added complications (World Health Organisation 2016c). *Salmonella* infection resulted in the second highest number of deaths and the greatest number of hospitalisations due to zoonoses in the EU in 2017 (European Food Safety Authority 2018). Non-typhoidal *Salmonella* was the most frequent cause of foodborne

pathogen outbreaks recorded in England and Wales from 1992-2008, with 1135 outbreaks (Gormley *et al.* 2010), and resulted in 24% of all outbreaks attributed to foodborne disease in 2017 across Europe (European Food Safety Authority 2018). It is reported that 1% of FBO caused by *Salmonella* in 2017 were strongly-linked to contaminated fresh produce (European Food Safety Authority 2018). *Salmonella* has been shown to survive and proliferate for 2-8 months in contaminated soil meaning any contamination event can have a considerable impact (Monaghan and Hutchinson 2010).

1.1.1.2 Campylobacter

Campylobacter is the most common cause of zoonosis in the UK (NHS 2015b) and Europe (European Food Safety Authority 2016), although only a small percentage of cases have been linked to food (Table 1). There are several species of *Campylobacter* that are pathogenic to humans, with most cases attributed to *Campylobacter jejuni* (European Food Safety Authority 2016). Symptoms of infection include diarrhoea, abdominal pain, fever, headache and nausea/vomiting. The incubation period of the disease is usually two to five days with symptoms lasting for three to six days, although it typically does not require treatment. It is predicted that one in 1000 cases of *Campylobacter* infection lead to Guillain-Barré syndrome (Nachamkin *et al.* 1998), an autoimmune disorder of the peripheral nervous system, that can lead to complications and death (World Health Organisation 2016a). Cases are predominantly attributed to the consumption of raw or undercooked poultry, however, models suggest that approximately 0.3% of fruit and vegetables are contaminated with *Campylobacter;* likely through contamination of soil or produce directly by animal faecal matter (Verhoeff-Bakkenes *et al.* 2011). *Campylobacter* can survive for 2-3 weeks in soil at ambient temperatures and for up to 2 months in cool, moist soil or water, increasing the risk that produce may become contaminated via direct contact (Monaghan and Hutchinson 2010). *Campylobacter* is the second most common bacteria associated with FBOs and a number of reports have identified a correlation between Campylobacteriosis and consumption of fresh produce (Mohammadpour *et al.* 2018).

1.1.1.3 Listeria monocytogenes

Clinical infection with *Listeria monocytogenes* is low incidence, with only 2,480 confirmed cases in the EU in 2017; however a high proportion of cases lead to hospitalisation and mortality as shown in Table 1 (European Food Safety Authority 2018). The symptoms associated with *Listeria* infection include fever, muscle aches and pains, chills, vomiting and diarrhoea. These symptoms do not usually require treatment and pass within a few days.

However, if the infection spreads into the blood or central nervous system of the individual it can lead to severe listeriosis. In such cases, the symptoms are more extreme and may include severe headaches, stiffness of the joints and neck, seizures and tremors. Infection can also, rarely, lead to development of meningitis and septicaemia (NHS 2015c). In pregnant women, listeriosis can lead to pregnancy and birth complications, resulting in miscarriage or stillbirth. Death occurs in approximately 20% of listeriosis cases and the number of recognised cases is on the rise (European Food Safety Authority 2016). There are few FBOs associated with *L. monocytogenes* (Table 1) all of which are associated with ready-to-eat foods, including fresh produce. *Listeria monocytogenes* is able to grow and survive at refrigeration temperatures, in soil and in many food matrices, making it a particular risk in the fresh and/or minimally processed food sectors (Gandhi and Chikindas 2007; McLaughlin *et al.* 2011).

1.1.1.4 Escherichia coli

Zoonotic infection resulting from *Escherichia coli* is primarily due to Shiga toxin-producing strains (STEC). Infection with STEC is relatively rare but results in high levels of mortality (Table 1). The incubation period of STEC is usually three to four days but can be as protracted as 14 days. Symptoms of infection include diarrhoea (bloody in about 50% of people), stomach cramps and occasionally fevers. These symptoms last for up to two weeks but usually do not require treatment.

In approximately 10% of cases infection with STEC leads to development of haemolytic uremic syndrome (HUS), which is a serious condition that can lead to haemolytic anaemia, low platelet counts, kidney failure and, ultimately, death. Children under five-years-old are at greatest risk (World Health Organisation 2016b). Treatment of STEC infection can increase the risk of HUS and therefore administration of antibiotics or anti-diarrhoea drugs is not recommended since it can prolong exposure to the toxin responsible for HUS (NHS 2015a). Shiga toxin-producing *E. coli* has been linked to several high profile outbreaks associated with fresh produce, including the sprout-associated outbreak in 2011 (Buchholz *et al.* 2011). Shiga toxin-producing *E. coli* can survive for up to two months in manure and the toxin is powerful, causing illness at levels of exposure lower than detectable using conventional screens (Monaghan and Hutchinson 2010; Tomás-Callejas *et al.* 2011). Several recent fresh produce-associated outbreaks that have been traced back to STEC-related agents have resulted in growing research interest in this pathogen.

1.1.1.5 Viruses

Viruses are the second highest cause of gastrointestinal infections in the UK. It has been estimated that up to 95% of clinical cases occur due to person to person transmission (O'Brien *et al.* 2016). Viruses are unable to reproduce outside their hosts, meaning food contamination events are most likely to happen at harvest or during post-harvest handling processes, when human contamination can inadvertently transmit high titres of virus particles. Viruses have been shown to survive on crops and/or in soil for at least a month at refrigeration temperatures (Monaghan and Hutchinson 2010) and may persist for longer than the shelf life of the product (Cook *et al.* 2016). There is currently no treatment for foodborne viral infection which makes prevention of over-riding importance (NHS 2016a, b).

Norovirus

Contraction of norovirus rapidly results in a suite of symptoms including diarrhoea, nausea, violent vomiting and mild fever which are short in duration. The incubation period is around 28 hours (Lee *et al.* 2013). Complications are rare due to the self-limiting nature of the virus, with mortality typical only in the extremely young, old and/or weak. The virus has been found to lead to asymptomatic infection in a high percentage of individuals, during which time individuals remain highly infectious (Robilotti *et al.* 2015). There is currently no vaccine available to prevent norovirus infection.

Rotavirus

Rotavirus infection leads to symptoms similar to norovirus, with the main clinical manifestation being diarrhoea. Children are most at risk since infection is usually asymptomatic in adults (Greenberg and Estes 2009). The virus is self-limiting but the symptoms can result in life-threatening dehydration in children; up to 10% of children who have contracted the virus end up in hospital and require rehydration (NHS 2017). A vaccine against rotavirus is available for young children, but it is not administered routinely in many countries (Greenberg and Estes 2009).

Hepatitis

Hepatitis is a cause of viral liver disease which can lead to mild to severe illness. Symptoms include tiredness, joint and muscle pain, fever, nausea or vomiting, jaundice and itchy skin although not all infections are symptomatic. Incubation time of the disease is around four weeks and symptoms last for up to two months, but it is usually self-limiting, and fatality is rare. Only hepatitis A and hepatitis E are associated with FBOs. In the case of hepatitis E,

approximately 70% of infections are asymptomatic (Van der Poel *et al.* 2018), whereas infection by hepatitis A is usually symptomatic in all bar young children (Pintó *et al.* 2010). The primary transmission route for foodborne-hepatitis is *via* faecal contamination of foods (commonly salads) and/or drinking water (Van der Poel *et al.* 2018). A vaccine is available for hepatitis A for those considered at high risk.

1.1.2 Sources of Contamination of Fresh Produce with Human pathogens

Contamination of fresh produce by human pathogens is most frequently associated with faecal contamination, due to sewerage issues, contamination of soil or water sources used to produce crops. Contamination can occur at any point in the supply chain, although evidence suggests most outbreaks are due to pre-harvest contamination of the produce (Barak *et al.* 2010), with studies observing that multiple processing lines can become contaminated with the same microbial contaminant from the field (Kim *et al.* 2016). The extent and routes of contamination are believed to vary from country to country as there are widely variable standards of hygiene during agronomy/irrigation, harvest, and storage dependent on the origin of the produce (Heaton and Jones 2008a).

In recent years, a significant focus for research has been the potential transmission vectors of human pathogens within the food system. Research has shown that flies can become contaminated with *E. coli* during brief exposure to contaminated apple surfaces and are capable of transmitting the bacteria from one product to another (Janisiewicz *et al.* 1999). It has also been shown that wild birds can disseminate *Campylobacter*, *Salmonella*, *Listeria* and *E. coli* O157 and contaminate crops in the field (Beuchat and Ryu 1997). Wastewater is another potential contamination source that is a focal point for much current research; irrigation with waste water is increasing due to water shortages and droughts worldwide and this is increasing the risks posed as waste water is a common reservoir for human pathogens and may lead to bioaccumulation of human pathogens in soil (Chang *et al.* 2013; Sallach *et al.* 2015). Principle contamination sources are summarised in Table 2.

Table 2. Pre-harvest and post-harvest sources of contamination of fresh produce with human pathogenic organisms

Pre-harvest Contamination Sources	Post-harvest Contamination Sources
Faeces	Faeces
Soil	Human handling
Manure	Harvesting equipment
Slurry	Transport containers / vehicles
Irrigation water	Wild and domestic animals
Fungicide / insecticide application	Insects
Air (dust)	Air (dust)
Wild and domestic animals	Wash and rinse water
Insects	Processing equipment
Human handling	Ice
Compost	Improper storage or packaging
	Cross-contamination from other foods
	Improper handling

Table adapted from Beuchat and Ryu (1997)

1.1.3 Prevention of Fresh Produce Contamination

Currently there are no acceptable methods of removing or inactivating human pathogens on fresh produce that do not affect the produce itself *via* the potential introduction of taints or residues (Sela Saldinger and Manulis-Sasson 2015). Post-harvest washing with oxidative agents, often with chlorine-based disinfectants, is often employed in the processing of fresh produce, but the use of chlorine for the reduction of viable human pathogens is of limited efficacy at the low concentrations generally employed in the industry (Beuchat and Ryu 1997) and has been linked to the formation of potentially carcinogenic compounds in wash water (Rico *et al.* 2007).

Due to the lack of acceptable decontamination methods, efforts and regulations are often focused on the prevention of contamination. The Codex Alimentarius Commission (Adopted 2003. Revision 2010 (new Annex III for Fresh Leafy Vegetables), 2012 (new Annex IV for Melons), 2013 (new Annex V for Berries)) outline key hygiene practices applicable to fresh fruit and vegetables and cover primary production through to packing and transportation, including hazard analysis critical control point (HACCP) analysis. Further guidelines published by the UK Department for Environment Food and Rural Affairs (2015) outline current restrictions on UK farmers with regard to spreading and storage of organic manures to prevent contamination of produce. Despite guidelines and controls to prevent contamination, a lack of time, knowledge and financial pressures can prevent their correct implementation. Adoption of high standards of hygiene during cultivation, harvesting and processing of crops destined for the food chain and the implementation of HACCP are essential for the proactive control of foodborne contaminants. Food safety cannot be based on end product testing alone (Zwietering *et al.* 2016).

Previous research has focused on ecological and microbiological factors that influence the risk of contamination by human pathogens and the survival of human pathogens on fresh produce. Maturity of the fruit, season and cultivar have been correlated with susceptibility to contamination by human pathogens (Barak *et al.* 2010; Chang *et al.* 2013). In addition, environmental factors can affect the survival of human pathogens associated with fresh produce in the field environment, such as exposure to UV light which can decrease their persistence (Wood *et al.* 2010). Mechanical damage of the produce may also affect the survival of human pathogens, potentially due to the increase in free available nutrients (Fatica and Schneider 2011; Deering *et al.* 2012). Research has also shown a negative

correlation between the diversity of the soil microbiota and survival of bacterial human pathogens (van Elsas *et al.* 2012), as well as correlations between specific bacteria and the survival of human pathogens (Wells and Butterfield 1997; Heaton and Jones 2008a).

1.1.3.1 Internalisation

A challenge to the prevention of foodborne disease caused by fresh produce is the potential for disease-causing organisms to internalise within the plant. Internalisation may confer a protective effect to human pathogens against multiple stressors including UV exposure, desiccation, temperature changes and nutrition changes (Heaton and Jones 2008a; Bartz *et al.* 2015). It also decreases the efficacy of decontamination as internalised organisms are protected within the structure of the produce (Gandhi *et al.* 2001; Warriner *et al.* 2003).

The ability of human pathogens to internalise within fresh produce is affected by a large number of factors; notably the species and genera of endophyte, crop type and age (Brandl and Amundson 2008; Golberg *et al.* 2011), soil type (Zhang *et al.* 2016b), water availability (Zhang *et al.* 2016b), and native endophytic and phytopathogen populations (Ge *et al.* 2014). A key risk factor for internalisation within the food supply chain is when the crop and bacteria are co-located in an aqueous environment, with an increased likelihood of internalisation if the water is of lower temperature than that of the crop (Buchanan *et al.* 1999). This is likely due to the increase in passive movement of contaminated water into the crop due to the presence of a temperature differential (Heaton and Jones 2008a). To account for this, current processing methods require any steps involving the washing of produce in water to maintain the water and produce at the same temperature. However, in the field, there may be differences in the temperature of crops and irrigation water which could lead to increased risk of internalisation (Burnett *et al.* 2000).

The evidence supporting the internalisation of pathogens is inconclusive and findings are often contradictory (Warriner and Namvar 2010). As a consequence, it is generally concluded that the risks of foodborne illness occurring as a result of the internalisation of pathogens is low (Monaghan *et al.* 2008).

1.1.3.2 Biofilm Formation

Biofilms are aggregations of bacteria within an extracellular matrix that can contain a single species of organism or, more frequently, a diverse range of organisms. The matrix can anneal to biological substances, such as foodstuffs or soil, as well as to hard non-natural

surfaces, such as equipment or storage surfaces (Galie *et al.* 2018), and formation is possible across a range of temperatures (Bonsaglia *et al.* 2014). The formation of biofilm generally results in the enhanced survival of bacteria on food and in the food production environment. In addition, biofilm formation leads to a greater resistance to antibiotics and decontamination using biocides (Russell 2003; Condell *et al.* 2012). This increased tolerance, with bacterial cells exhibiting 10 to 1,000 times less susceptibility to specific antimicrobial agents within a biofilm (Balcazar *et al.* 2015), is due to numerous mechanisms including failure of the biocide to penetrate the biofilm and increased stress response of cells inside a biofilm (Olsen 2015). Low concentrations of antimicrobials or biocides may also lead to an increase in horizontal gene transfer of AMR associated genes due to an increase in the stress response of members of the biofilm, aided by the close proximity of cells within a biofilm environment, providing optimum conditions for conjugation (Balcazar *et al.* 2015). Human pathogens isolated from fresh produce have been shown to be able to form biofilms (Amrutha *et al.* 2017), and human pathogens have been found as part of biofilms both on fresh produce and within the fresh produce processing environment (Galie *et al.* 2018). The increased persistence of bacteria in biofilms and the potential for increased biocide resistance and transfer of AMR genes make food associated biofilms a key risk within the fresh produce supply chain.

1.1.4 Antimicrobial and Biocide Resistance

Antimicrobial resistance (AMR) develops when bacteria adapt to grow in the presence of antimicrobial compounds or elements, often through genetic changes, and leads to the acquisition of resistance to the antimicrobial compound. Antimicrobial resistance can occur through mutation of genes already present within bacteria, or through acquisition of new genetic material through horizontal gene transfer. There are multiple mechanisms of horizontal gene transfer including transformation, transduction and conjugation. Transformation is the uptake of naked DNA into a bacterial cell from the surroundings, and potentially allows for the transfer of genetic material from distantly related cells. Transduction is the transfer of genetic elements from one bacterium to another *via* a bacteriophage. The phage can carry the genetic element into their host cell with them upon infection. Due to bacteriophage's specificity to certain species or genera this mechanism is limited to the transfer of genetic material to related species. Conjugation is the transfer of genetic material *via* direct contact between a donor cell and the recipient bacteria, therefore

is more likely to occur in cells which are in close proximity, for example in a biofilm. As there is less likelihood of two different species being able to create a direct contact this mechanism is most likely to allow transfer between bacteria of the same species or genera (Gyles and Boerlin 2014).

The mechanisms of action of AMR broadly fall into four categories: drug efflux, where the antimicrobial is actively pumped out of the cell thereby not allowing toxic levels to build up in the cell (Levy 2002); decreased cell membrane or wall permeability, where microbial cell structure is altered and therefore antimicrobial compounds cannot enter and reach toxic levels (Delcour 2009); target overexpression, modification or protection, where the target of the antimicrobial compound is altered in some way to allow for its function even in the presence of the antimicrobial compound (Adu-Oppong *et al.* 2016) and drug inactivation, where the cell gains the ability to render the antimicrobial compound inactive, for example through the enzymatic destruction of the active site of the compound (Lakaye *et al.* 1999). AMR genes often confer resistance to multiple antimicrobials due to their mechanism of action (Szmolka and Nagy 2013). The transfer of AMR genes has been demonstrated experimentally within the fresh produce environment and many studies report AMR within human pathogens isolated from fresh produce (de Vasconcelos Byrne *et al.* 2016; Zhang *et al.* 2016a). There is also a high prevalence of AMR genes and microbes with AMR genes within the microbiota of fresh produce and soil (Pedroso *et al.* 2013; Rolain 2013). The high microbial density in these microbiomes favours the transfer of AMR genes (Aarts and Margolles 2015). The application of manure to soil also leads to increased numbers of AMR genes, both free and within members of the microbiome, creating a reservoir of AMR genes that can pass to human pathogens (Zhu *et al.* 2016). Herbicides including Glyphosate, have also been linked to driving antibiotic resistance in human pathogens (Kurenbach *et al.* 2015).

The presence of AMR genes may confer advantages to the host in terms of, for example, resistance to biocides. Biocides are compounds used to inactivate bacteria and are widely used in the food production chain to decrease the likelihood of contamination of foods, feeds and drinks with human pathogens. Due to the overlap in some mechanisms of resistance to antibiotics and biocides there is the potential for the evolution of cross-resistance to both types of compounds. This can occur when the biocide and antibiotic have the same target or the same transport mechanism into the cell (Condell *et al.* 2012). Biocides are typically lethal to their target, often after a single application, and generally have

multiple targets (Wales and Davies 2015). Therefore, the main mechanisms utilised to confer biocide resistance are usually drug efflux or decreased cell membrane or wall permeability (Russell 2003). Target site mutations are much rarer in biocide resistance than AMR (Poole 2002). Biocides are often deployed at much higher concentrations than many bacterial resistance mechanisms can cope with, therefore it is more difficult for resistance to emerge (Condell *et al.* 2012). Where resistance genes are already present the addition of biocides may drive the transfer of these genes between bacteria, notably in biofilms which will decrease the concentration of the biocide in contact with the bacteria, thereby allowing resistance mechanisms to be selected for.

AMR is a growing problem, with many initiatives such as the Global Action Plan on Antimicrobial Resistance (World Health Organisation 2015a), being developed to try and prevent the world entering a "post antibiotic era". The consequences of antimicrobial resistance include limited treatment options, longer and more severe illness, greater mortality and increased costs.

1.1.5 Bacteriophages

The use of bacteriophages to eliminate bacterial pathogens is increasingly being studied as an alternative to antibiotic or biocide use. Bacteriophages are obligate intracellular parasites of bacteria that multiply through use of the host biosynthesis machinery and fall under the classification of viruses. They are specific to bacteria, although many of the molecular interactions between phage and their bacterial hosts remain largely unexplored. The observable effects of bacteriophages were first reported by Hankin (1896), who observed the antibacterial effect of a component of the Ganges and Jumna rivers in India which could pass through a fine filter and retain its activity. It has since been found that bacteriophages are ubiquitous in the natural environment, notably in soil and water (Chibani-Chennoufi *et al.* 2004). Bacteriophages play a key role in the transfer of genetic elements between prokaryotes. Segments of host DNA can become encapsulated within the bacteriophage and, upon attachment of the phage to a new host, transferred to the new bacteria which may then incorporate the DNA into its genome. This makes them an important vector for resistance associated genes. Bacteriophages are specific to their host, although the host range varies, with some being specific to a single strain of bacteria, whilst others are capable of parasitizing several members of a bacterial family (Adams 1959). They have no known effects on humans, animals, plants or non-host bacteria. Their ubiquity also indicates they

are safe for contact with the environment and humans. The FDA approved the first bacteriophage treatment as a food additive in 2006 (U.S. Food and Drug Administration 2014). The use of bacteriophages in the food industry, as well as for other commercial applications, is an area that is still under development and the subject of much commercially oriented research.

1.2 Current methodologies for detection of plant associated human pathogens

Several methodologies may be employed to detect and identify foodborne pathogens, or indicators of faecal contamination, on fresh produce. Enrichment and plating onto specific agar-based solid medium to detect the pathogen(s) of interest is the basis for most International Standards. Enrichment and plating rarely identifies a single target species and therefore must be repeated with other selective agars to enrich the bacterium of interest. Due to the low prevalence and titre of human pathogens on fresh produce, indicator organisms are often utilised to indicate the presence of a contamination event or poor hygiene practice (Health Protection Agency (now Public Health England) 2009). These indicator organisms are commonly present in higher numbers than the potentially-associated pathogens and are generally easier to culture and identify. Health Protection Agency (now Public Health England) (2009) guidelines for ready-to-eat foods list Enterobacteriaceae, *Escherichia coli* and *Listeria* species as indicator organisms in fresh produce and identifies levels that raise concern. Plating methods are an important, if not vital, tool as they are presently the only methods capable of identifying microbes that can grow and replicate, therefore are a potential risk, without confusion arising from inactivated or damaged cells or free DNA.

One disadvantage of culture-based methods is that they are labour-intensive; time consuming; require replication; and when conducted in bulk to service commercial demands, relatively expensive. There is also the potential for artefacts resulting from competition on plates, where non-target organisms outcompete the targeted microorganisms for space on the plate thus impacting on the recovery of the targeted organism(s). For example *Listeria monocytogenes* usually occurs at low levels and often in similar environments as species that are known to outcompete it in culture e.g. other *Listeria* spp. (Keys *et al.* 2013; Dailey *et al.* 2015). False negatives for *Listeria monocytogenes* are also frequently encountered as *Listeria* spp. have indistinguishable colony morphologies and only five colonies per plate are required to be tested to determine species to meet

UKAS/ISO-testing requirements (Oravcová *et al.* 2007). Additional issues with culture-based approaches include the time delay between culture and identification, which can be several days for most common pathogens such as *L. monocytogenes* and *Salmonella*, and the fact that culture must be combined with further tests (e.g. serological assays, biochemical test strips, pulsed-field gel electrophoresis (PFGE), flow cytometry and polymerase chain reaction (PCR)) to confirm species identification and sub-typing of isolates. To gain more detailed information on the relatedness of strains for source tracking and outbreak control, further approaches such as multi-locus sequence typing (MLST) must be performed.

To counter these disadvantages molecular methodologies, such as PCR, real time PCR and Loop-Mediated Isothermal Amplification (LAMP), are increasingly being applied for routine microbiological assessments and are standard for viruses. Next Generation Sequencing (NGS) approaches are also attracting much interest as a possible means to explore and profile the microbiome associated with fresh produce.

1.3 Next Generation Sequencing

1.3.1 Development of NGS Technologies

DNA sequencing is a methodology for determining the sequence of bases within the genetic material of an organism. The first method to be widely applied was the 'chain-termination' method described by Sanger and colleagues (1977). This method utilises adapted deoxyribonucleotides (dNTPs), which prevent amplification, in conjunction with unmodified dNTPs and amplification using PCR, to lead to a termination of amplification at every point in the DNA chain (Figure 1). The mixed DNA with terminations corresponding to every nucleotide is then visualised to read the sequence of the DNA. This visualisation was initially undertaken on agarose gels, but the invention of automated DNA sequencers led to easier visualisation and automation (Smith *et al.* 1986). The technology was costly, in both money and laboratory hours, inhibiting the adoption of DNA sequencing in any routine manner, until the advent of next generation sequencing (NGS) technologies.

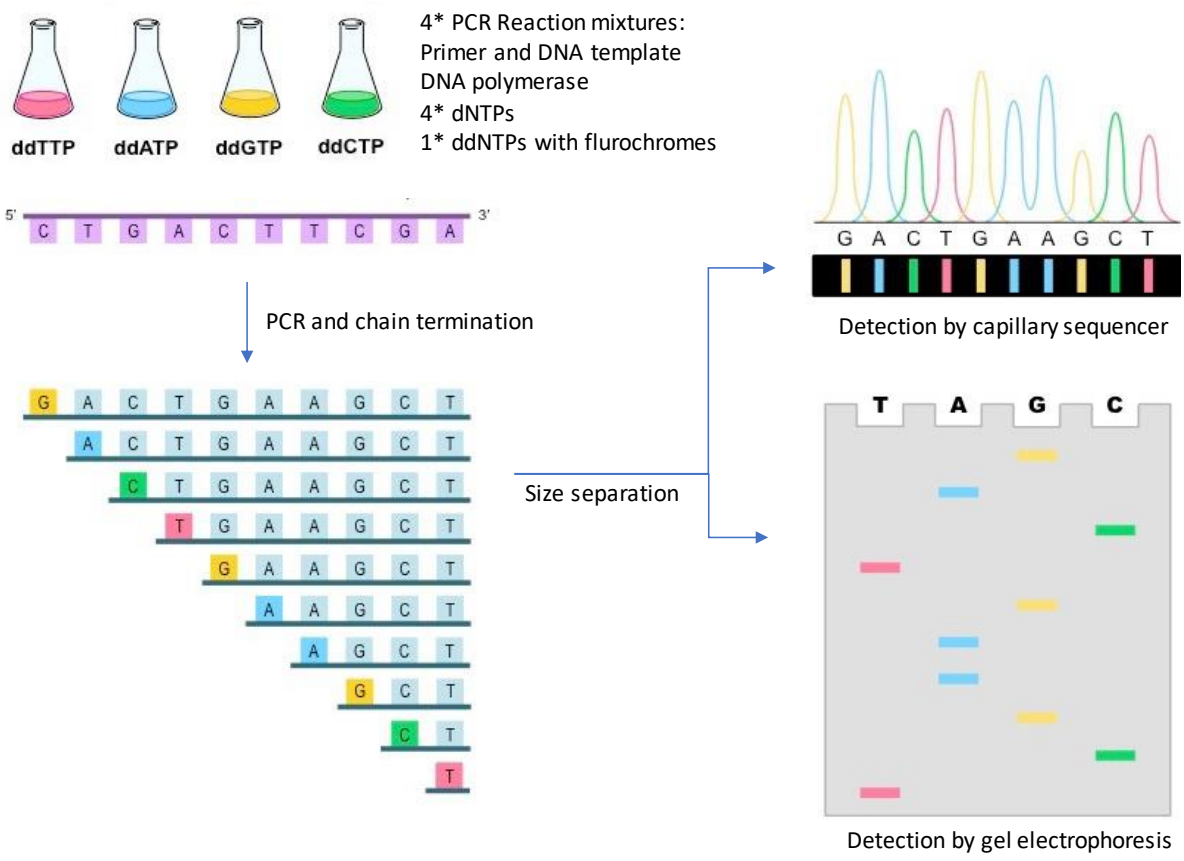


Figure 1. Outline of Sanger sequencing methods showing use of ddNTPs to terminate elongation, and detection of DNA sequence using capillary electrophoresis and gel electrophoresis. Figure adapted from Karki (2017).

The first widely adopted NGS platform was the 454 Pyrosequencer. Released in 2005, it was the first commercial sequencer to produce reads of a useable quality and length. Utilising similar methods to Sanger sequencing, the 454 used sequence by synthesis technology and light signals to take a snapshot of each base added and infer DNA sequence. The normal read length of the 454 was between 400 and 500 bases, although subsequent developments allowed for read lengths of 500-800 bp, with approximately one million sequences produced and a run time of four hours (Margulies *et al.* 2005). The 454 was the first machine that allowed for parallelisation of sequencing, dramatically increasing the amount of DNA sequence that could be generated and decreasing the cost of sequencing per base (Heather and Chain 2016). Following the success of the 454 many other platforms were developed, some focusing on increasing the read length, others on the accuracy and cost of sequence per base.

Another key NGS platform was the Ion Torrent, originally produced by Life Technologies. This machine originally used similar methods to the 454, but instead of using luminescence to visualise the sequence it measured pH changes caused by the release of protons (H^+ ions) during polymerisation. This allowed machine prices to be reduced since the expensive optical requirements of the 454 were circumnavigated.

Arguably the most important NGS development was the introduction of Solexa methodologies used by Illumina in their MiSeq and later the HiSeq. Briefly, the Solexa method requires the immobilisation of DNA on a plate known as a flow cell, and amplification of this DNA (known as bridge-PCR) to create clusters of identical DNA, which is then sequenced by washing labelled dNTPs over the flow cell. These labelled dNTPs anneal and are excited by lasers leading to fluorescence, thereby allowing identification of the base (Illumina 2010). This can be duplicated for both forward and reverse strands of the DNA to create paired-end reads. The MiSeq platform is capable of sequencing DNA fragments up to 600 bp (through two 300 bp reads), and is capable of generating far more data per run than the 454 or Ion Torrent machines, but in order to obtain these read lengths it requires a greater run time than other machines, taking 56 hrs, although shorter read lengths can be achieved in less time. The workflow for sample preparation for the MiSeq is also quicker and less prone to errors than the 454 and Ion Torrent methodologies. The ease of use and quality

of the data resulting has resulted in Illumina becoming the market leader in current sequencing technologies.

The methods described above give the sequence of a consensus of PCR product, or amplified fragment of DNA. The amplification of DNA may lead to errors during sequencing *via* the incorporation of incorrect bases in the amplification stages. In addition, by taking a consensus, some of the diversity of single molecules within the sample may be lost. The development of single molecule sequencing (SMS) technologies has allowed the direct sequencing of a single molecule of unamplified DNA. The first platform to use such SMS technology was the single molecule, real time (SMRT) platform from PacBio. Sequencing is performed using an array of micro-fabricated nanostructures called zero-mode waveguides (ZMWs) that allow light to exclusively illuminate the bottom of a well in which the DNA and polymerase complex is immobilised. Modified dNTPs are then used to extend the DNA strand allowing detection of the base sequence in real time. This methodology allows for the sequencing of fragments of DNA from 250 bp to 40 kb in 0.5 to 4 hours (Pacific Biosciences 2015), which is particularly useful when sequencing whole genomes. The data produced is less accurate than that of Solexa methods and therefore is less applicable where single nucleotide differences are of importance, although laboratory methods such as nucleotide circularisation can allow for increased accuracy.

Another SMS method is nanopore sequencing. The first nanopore sequencer to be commercially available was the MinION, followed by the larger PromethION and GridION, marketed by Oxford Nanopore. These sequencers work through the creation of single stranded DNA (ssDNA) which is passed through a protein nanopore immobilised within an impermeable “membrane”. The presence of the ssDNA within the nanopore alters the current across the membrane, with a different change in current occurring depending on which bases are present within the pore. This nanopore then drives the DNA across the membrane one base at a time allowing for the changes in current to be read and therefore the sequence of the DNA to be calculated. The Oxford Nanopore sequencers give relatively long read length (currently believed to be limited only by current extraction technologies) and are both cheaper and faster than other sequencing technologies, although there are currently significant error rates (Laver *et al.* 2015).

1.3.2 Current NGS Techniques and Applications

Continuing developments in NGS technologies have allowed for higher throughput of samples with greater yield of quality data. With this has come a sharp reduction in the costs of sequencing per mega base of data (Wetterstrand 2019) allowing for a further increase in the use of sequencing technologies. The decreased cost platform also allows the expansion of applications for DNA sequencing, and opens-up opportunities for the routine use of NGS in some sectors where it would previously never have been considered feasible (see Table 3).

Table 3. Description of techniques used in NGS and examples of their applications

Technique	Description	Example applications
Whole genome sequencing	Total genome of a single organism	Subtyping, characterisation of new or emerging pathogens
Metagenomics	Total DNA for the whole microbiome of a single sample	Environmental genetics
Amplicon sequencing / Metabarcoding	Focused DNA (one target gene) for whole microbiome for a single sample	Biodiversity assessment
Metatranscriptomics	Total RNA for whole microbiome for a single sample	Gene activity, gene expression, differential gene expression

Following on from the success of projects such as the Human Genome Project, whole genome sequencing (WGS) has been widely utilised in applications in human health, including assessment of bacterial genomes for outbreak and phenotypic analysis (Gilchrist *et al.* 2015; Shen *et al.* 2015; Ellington *et al.* 2017). Frequently utilised for the identification, subtyping and characterisation of human pathogens (Black *et al.* 2015), WGS is now routinely applied in food microbiology (Gilmour *et al.* 2010; Allard *et al.* 2012; Franz *et al.* 2014; Joensen *et al.* 2014). The approach potentially allows the improved discrimination of similar species over traditional methodologies such as PFGE and MLST (Bergholz *et al.* 2014), facilitating application in surveillance programmes, for example the US FDA's GenomeTrakr, the goal of which is to provide identification of outbreaks and source tracking of causal strains (U.S. Food and Drug Administration 2019). In addition, WGS data provides a wealth of genetic information which can be used to screen for genes of interest, for example those associated with virulence or antimicrobial resistance (Critzer and Doyle 2010; Forsberg *et al.* 2012).

Next generation sequencing also allows for examination of the microbiome associated with foods, feeds, drinks, soils and water in greater depth than conventional targeted approaches as well as facilitating the identification of non-culturable elements of the microbiome (Trček et al. 2016; Andersson et al. 2008). Assessment of the total microbiome is a widely applied NGS approach used in food microbiology (Ercolini *et al.* 2011; Guzzon *et al.* 2014; Trček *et al.* 2016). There are several methodologies that can be employed; metabarcoding, also known as amplicon sequencing, metagenomics and metatranscriptomics. Key uses of these methodologies are outlined in Table 3.

Metabarcoding involves the amplification and sequencing of a gene universal to the target population (an amplicon), for example the 16S rRNA gene for bacteria or ITS gene for fungal populations. This target gene must have a variable region, used for identification, flanked by highly conserved regions, which enables annealing of the primers. By focusing on a small target gene, it allows for differentiation but requires less sequencing per sample than techniques that are totally non-selective. This is the oldest and most frequently utilised method for assessing the microbiome due to its lower cost (Johnson *et al.* 2019). Due to this pipelines and databases for analysis of this data are more mature, with most users of this data choosing pipelines, such as QIIME (Caporaso *et al.* 2010) or mothur (Schloss *et al.* 2009), in conjunction with databases such as SILVA (Quast *et al.* 2013) or GreenGenes (DeSantis *et al.* 2006). Additionally, tools to allow users without access to large computing clusters to analyse data are also available allowing for greater use of these technologies, one such example being MG-RAST (Meyer *et al.* 2008). One issue with metabarcoding is that, on the most commonly-used platforms, the length of the amplicon is too short to achieve accurate identification below genera level for many widely used barcodes including the 16S rRNA gene amplicon employed for bacterial microbiome analysis (Janda and Abbott 2007). This may change in the future with the utilisation of longer read technologies, such as the MinION (Benítez-Páez *et al.* 2016), but currently the error rate prohibits this application of nanopore technology. Another limitation is the potential for bias introduced by the amplification of the amplicon (Kennedy *et al.* 2014).

Metagenomics involves direct sequencing of the total DNA within a sample and yields information on the total microbiome of the sample through direct isolation and sequencing of nucleic acids. Unlike metabarcoding this is a non-selective approach and therefore will be less likely to lead to PCR biases associated with some barcodes and predominantly allows for

species level identity. This method also facilitates the examination of the total microbiome, including bacteria, fungi and DNA viruses, although it is unable to detect RNA viruses. This method has been increasingly utilised to examine the microbiome, and bioinformatics tools being developed to allow for this, such as Kraken (Wood and Salzberg 2014) or MetaPhlAn2 (Truong *et al.* 2015), and online tools are additionally becoming available such as through the MG-RAST system. These bioinformatic methods have not been widely used or validated and as yet there is no standard methods or databases that are used in the analysis of these data.

Metatranscriptomics involves sequencing of the total RNA from a sample and thereby can be used to study gene expression and activity and, in conjunction with WGS, is a useful tool to explore alternative splicing patterns and differential gene expression (Wang *et al.* 2009). Recently, it has been applied to the study of the microbiome, notably for viruses (Bashiardes *et al.* 2016). As with metagenomics, the approach is non-selective and allows for the examination of the total microbiome, including bacteria, fungi, RNA viruses and DNA viruses that are being actively transcribed within their host. The approach is unable to detect DNA viruses that are not actively replicating. Due to this method again being applied in the field of the study of the microbiome, it too suffers from a lack of standardised bioinformatic methods and databases. The methods used to analyse this data are often adapted metagenomics workflows such as using Kraken or MG-RAST, or programs such as BLAST (Camacho *et al.* 2009) which requires high computing power and time costs. These methods still require validation of their use in microbiome studies, to allow for accurate and efficient processing of data. By sequencing RNA, metatranscriptomics also yields potentially valuable information on patterns of gene expression in the microbiome (Maurice *et al.* 2013), another recent utilisation of this data, which as yet has minimal standardised or validated analysis protocols.

Application of NGS approaches to food, feeds and drinks may allow the proactive screening of samples for human pathogens. This would allow samples to be screened for multiple agents without the need for multiple tests. A significant percentage of foodborne illness arises as a result of unidentified agents and the use of non-targeted approaches such as NGS may also allow for the identification of novel, rare or non-culturable causal agents of foodborne disease (Scallan *et al.* 2011). The main issue with using NGS as a screening tool is that, as with all molecular techniques, these techniques struggle to distinguish between

living and dead cells (Bergholz *et al.* 2014), and studies have shown that when NGS is compared to CFU counts there is a significant difference in the results, attributed to the presence of dead cells (de Boer *et al.* 2015). This presently limits the use of NGS to an initial screening tool and follow-on tests must be performed to determine viability. In microbiome studies, where information on the community composition and ratios of each is most important, research has shown that the profiles of dominant taxa found using NGS is comparable to results found through culturing approaches (Jackson et al. 2015). Directed microbiome studies may help our understanding of the way in which the microbiome impacts on food spoilage and safety , in turn leading to improved shelf life, fewer costly recalls, and decrease of waste (Jackson *et al.* 2015). Further research into the fresh produce microbiome may also lead to the development of novel biocontrol agents or greater understanding of risk, for example the discovery of additional indicator organisms (Wall *et al.* 2015). Furthermore, understanding the microbiome may also allow for identification of sources of contamination within the fresh produce supply chain as different sources of contamination may impart different microbial profiles on the produce (Newton *et al.* 2013).

1.3.3 NGS Targeting and Enrichment Techniques

A drawback of NGS in the context of food microbiology is that many techniques have a large percentage of the data associated with the plant matrix, not the microbiome (Aw *et al.* 2016). Therefore, methods need to be employed to enrich the microbiome, or deplete the nucleic acid associated with the matrix.

Enrichment of the samples *via* the culturing of the sample prior to extraction may increase the proportion of the data associated with the microbiome but also significantly affects the microbiome (Rosimin *et al.* 2016; Hyeon *et al.* 2017). Culturing will additionally decrease the likelihood of detecting certain microorganisms, such as viruses or viable but non-culturable organisms (Highmore *et al.* 2018) negating the benefits of using a non-targeted approach such as NGS. Therefore, DNA based or post-extraction enrichment methods may be more suitable for use with NGS. There are several kits commercially available to enrich microbial nucleic acids, for example from Illumina, New England Biolabs or Qiagen. These predominantly use molecular probes to target either the fraction of the microbiome that is to be retained, for example poly-A capture, which targets the poly-A tail section of a virus, or targets the fraction to be removed, for example ribosomal depletion, that targets ribosomal nucleic acids leaving behind a supernatant containing non-ribosomal nucleic acids. There are

many probes available, although due to the predominance of research in human or mouse models, many of these kits are focused on removing the nucleic acids associated with these hosts. There are fewer kits available to target plant or food microbiomes, although the aforementioned poly-A capture and ribosomal depletion are two such methods. Additionally, many of these kits target RNA and therefore are only viable for use in conjunction with metatranscriptomics, not 16S or metagenomic methods. Although more research has been done recently into alternative depletion methods (Sun and Zu 2015; Lee *et al.* 2019; Song and Xie 2020), none are as yet commercially available and therefore have a substantial cost and time implication to them. Their lack of commercial availability also means a lack of validation and standardisation, ruling them out as candidates for routine and standardised screening methods.

1.4 Project Aims

This PhD focusses on the potential of next generation sequencing (NGS) to detect and characterise foodborne pathogens and elucidate the microbiome and potential influences on the survival and transmission of human pathogens within the fresh produce supply-chain.

The aims of this PhD were to:

(i) Develop laboratory and data analysis protocols that enable the identification of the microbiome of fresh produce and identify the limits of detection of these methods for human pathogens.
(ii) Analyse the fresh produce microbiome from samples obtained from the food supply chain and examine for correlations with microbiological data.
(iii) Use phenotypic and genotypic methods to characterise the resistome, virulome, and biofilm forming ability and assess the phylogenetic identity and gene content of these isolates compared to 80 isolates of meat and clinical origin to identify signatures of fresh produce contaminating *L. monocytogenes.*
(iv) Assess the incidence of AMR-associated genes in foodborne microbes.

Chapter 2. Development of MiSeq approaches for the detection of the human pathogens within the fresh produce microbiome

2.1 Introduction

Traditional techniques of pathogen detection on fresh produce allow for the theoretical detection of low levels of contamination on produce. The gold-standard approach, used as part of many standard methods for bacterial detection, requires plating onto selective and semi-selective media. This has a theoretical limit of detection (LoD) of 1 cfu (Bell *et al.* 2016), and also delivers valuable information on viability. Culture-based techniques are however limited, in that they only allow for the detection of organisms that are culturable – often a small fraction of the microbiome. Moreover, recent research has shown that many foodborne bacteria, including some potential pathogens, have viable but non-culturable (VBNC) forms, which have lost the ability to grow on media but remain infectious (Ayrapetyan and Oliver 2016). In addition, fully effective culturing techniques are not available for commonly occurring foodborne viruses such as norovirus. These limitations mean that molecular based methodologies are increasingly being utilised to screen food for human pathogens and spoilage organisms. The LoD for these techniques is often less sensitive than culture-based approaches although Loop Mediated Isothermal Amplification (LAMP) has a reported LoD of 1.3-28 targets/reaction (Domesle *et al.* 2018), and real time PCR a theoretical limit of 10 targets/reaction (Bell *et al.* 2016). Real time-PCR has proven a major advancement in the detection and surveillance of viruses, however obtaining information on the origin or relatedness of virus strains from different sources cannot always be achieved due to the generally small amplicon sizes produced yielding limited sequence information. Next generation sequencing (NGS) theoretically allows for the non-targeted detection of multiple spoilage agents and pathogens. As with other DNA based methods it allows for the detection of VBNC organisms and non-culturable organisms such as viruses. NGS is increasingly being utilised in research in food microbiology, with most studies focusing on whole genome sequencing (WGS) of bacterial isolates (Hyeon *et al.* 2017), and the exploration of the microbiome associated with specific commodities (Ottesen *et al.* 2013; Yi *et al.* 2017). Amplicon sequencing, for example of the 16S rRNA gene for bacteria, is currently the most widely applied technique for microbiome analysis due to its lower cost, development stage and standardised analytical pipelines (Cao *et al.* 2017). These

pipelines theoretically produce accurate and reproducible results, thereby allowing comparison between studies (Thompson *et al.* 2017; Bolyen *et al.* 2019).

The use of metatranscriptomics for interrogation of the microbiome is an emerging field and currently few tools are available to analyse data relating to the total microbiome, with those that are having been shown to produce variable results (Bashiardes *et al.* 2016). The use of metatranscriptomics for the interrogation of the microbiome has advantages, as it potentially allows for the WGS of members of the microbiome, including viruses. One of the key issues is contamination of the microbiome with host RNA; therefore, to obtain more data associated with the microbiome it is necessary to enrich the RNA associated with the microbiome. There are several methods for purifying or enriching this fraction. One example is poly-A capture. This allows for the purification of virus RNA through binding of the poly-A tail section of the virus allowing for removal of non-poly-A RNA. Another method is ribosomal depletion. This degrades ribosomal RNA leaving behind other RNAs, including viral RNA. Previous research has compared the efficacy of; multiple extraction methods (Hang *et al.* 2014; Fouhy *et al.* 2016), primer sets, PCR conditions (Ahn *et al.* 2012; Fouhy *et al.* 2016), and sequencing platforms (Caporaso *et al.* 2012; Quail *et al.* 2012; Allali *et al.* 2017), but very few have focused on the effects of different methodologies within the context of the fresh produce supply chain. In addition, scant attention has been paid to the effect of molecular enrichment methods on the microbiome or identified the LoD applicable to these methods. Those studies which have examined the LoD (Frey *et al.* 2014; de Boer *et al.* 2015), utilise outdated approaches or have used notably microbiologically sterile matrices which are less applicable to most real-world situations.

This study compared two methods of enrichment, poly-A capture and ribosomal depletion, on samples testing positive for norovirus using real time-PCR, followed by metagenomics analysis, to assess methodology for the detection of foodborne human pathogenic organisms. Findings were then applied to establish the limits of detection of current sequencing technologies using Illumina MiSeq and bioinformatic methods.

The aims of this study were to:

(i) Examine the effect of enrichment methodology (polyA capture vs ribosomal depletion) on the ability to detect viruses in metatranscriptomics studies using the ScriptSeq kit.

(ii) Identify and compare the limit of detection of metatranscriptomics (using NEBNext or ScriptSeq kits), and 16S rRNA gene amplicon sequencing (using primers to amplify the V4 region of the rRNA gene), on the MiSeq platform, via the analysis of *Salmonella* and MS2 phage spiked into a lettuce homogenate background.

(iii) Ascertain the impacts of different bioinformatic approaches on the assigned microbiome.

(iv) Use bioinformatic techniques to create a mock community to assess the effect of the differing read length and error profiles of current sequencing platforms on the LoD.

2.2 Methods

2.2.1 Comparison of Enrichment Techniques

2.2.1.1 Sequencing

Twelve RNA samples extracted from lettuce and raspberries were taken from an archive at Fera. They had been extracted following ISO Technical Specification 15216-1:2013, and contained mengo virus as a spiked internal positive control and had been tested for norovirus as part of a previous study (Cook *et al.* 2019). RNA was quantified using the Qubit® RNA HS Assay Kit (Invitrogen, Thermo Fisher Scientific, Carlsbad, United States) following the manufacturer's instructions (Life Technologies 2015b). Briefly, a working solution was made up by diluting Qubit RNA HS Reagent 1:200 in Qubit RNA HS Buffer in a clean appropriately sized tube. This was then vortexed and 190 µl added to two clean, low absorbance, 0.8 ml tubes and 10 µl of the appropriate standard added to each tube. For samples, 2 µl of the RNA extract was added to 198 µl of working solution. This was then gently vortexed and incubated at room temperature for 2 mins before reading on a Qubit V2 (Invitrogen, Thermo Fisher Scientific, Carlsbad, United States). Each sample was then split into two equal volumes, and one replicate of each was directed into one of two workflows, i) ribosomal depletion using the ScriptSeq™ Complete Kit for Plant Leaf (Illumina, San Diego, United States) following manufacturer's instructions (Epicentre 2013), or ii) poly-A capture using the NEBNext® Poly(A) mRNA Magnetic Isolation Module (New England Biolabs, Ipswich, United States) following manufacturer's instructions (New England Biolabs 2018). The processing of samples is outlined in Figure 2.

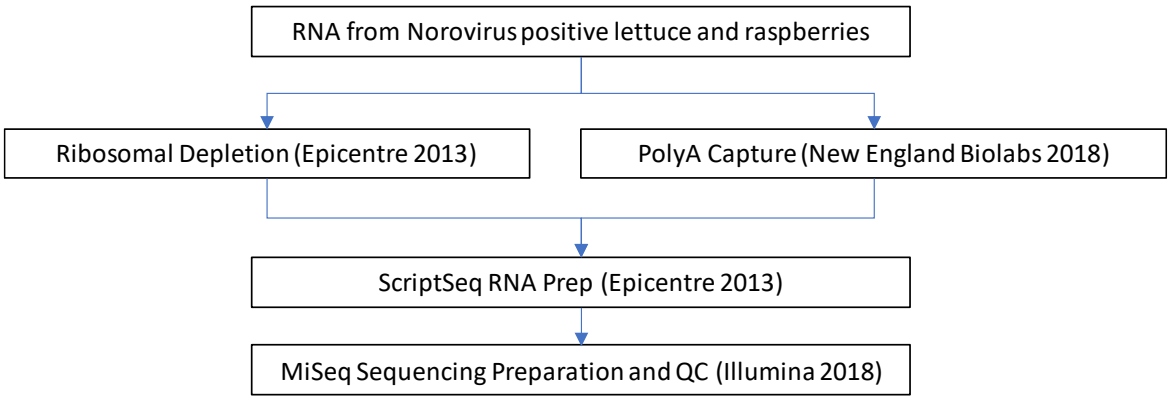


Figure 2. Flow chart showing methodologies employed for norovirus positive samples to compare two enrichment techniques, ribosomal depletion and polyA capture, from extracted RNA to sequencing.

Samples in workflow i were diluted with MGW to create a total volume of 28 µl. The samples were added to a 96 well PCR plate and 8 µl of rRNA Removal Solution and 4 µl Reaction Buffer added. The mixture was heated in a thermocycler at 68 °C for 5 minutes and then 22 °C for 5 minutes before being removed to RT. Magnetic beads were washed twice, resuspended and RiboGuard RNase Inhibitor added before 65 µl of beads was added to each sample and incubated at RT for 5 minutes. The mixture was then heated in a thermocycler at 50 °C for 5 minutes, before being removed to RT and placed on the magnetic rack and left for 2 minutes for the supernatant to clear. The supernatant containing the ribosomal depleted RNA was then transferred to a new 96 well PCR plate. The samples were cleaned by addition of 160 µl RNAclean XP (Beckman Coulter) and incubated at RT for 15 minutes. The mixture was placed onto a magnetic rack and left for 2 minutes for the supernatant to clear. The supernatant was removed, and the beads were washed twice in 80% Ethanol, left to air dry for 10 minutes before being resuspended in 12 µl molecular biological grade water (MBGW), left for 5 minutes at RT to elute the RNA from the beads, and placed on the magnetic rack and left for 2 minutes for the supernatant to clear. The supernatant containing the ribosomal depleted RNA was then transferred to a new 96 well PCR plate ready for processing using the Illumina ScriptSeq kit.

Samples in workflow ii were diluted with MGW to create a total volume of 50 µl. NEBNext Magnetic Oligo d(T)25 Beads were washed twice and mixed with the RNA sample in a 96 well PCR plate. The mixture was heated in a thermocycler at 65 °C for 5 minutes and then the temperature decreased to 4 °C, to denature the RNA and facilitate binding of the poly-A-

RNA to the beads, before being removed to RT, resuspended by pipetting gently, incubated at RT for 5 minutes, resuspended again and incubated at RT for a further 5 minutes. The mixture was placed onto a magnetic rack and left for 2 minutes for the supernatant to clear. The supernatant was removed, and beads were washed twice in Wash Buffer and then resuspended in 50 µl Tris buffer. The sample was heated in a thermocycler at 80 °C for 2 minutes, then the temperature decreased to 25 °C, to elute the poly-A RNA from the beads, before being removed to RT. 50 µl RNA Binding Buffer was added, to allow the polyA-RNA to rebind the same beads and incubated at RT for 5 minutes. The mixture was placed onto a magnetic rack and left for 2 minutes for the supernatant to clear. The beads were washed twice in Wash Buffer and then resuspended in 17 µl Tris buffer. The sample was heated in a thermocycler at 80 °C for 2 minutes, then the temperature decreased to 25 °C, to elute the poly-A RNA from the beads, before being removed to RT and placed on the magnetic rack and left for 2 minutes for the supernatant to clear. The supernatant containing the polyA RNA was then transferred into a new 96 well plate ready for processing using the Illumina ScriptSeq kit.

The samples then underwent ScriptSeq RNA preparation for sequencing. For sequencing using the ScriptSeq kit, 9 µl of each sample from both workflows had 1 µl RNA fragmentation Solution and 2 µl synthesis primer added to them and were heated in a thermocycler at 85 °C for 2 minutes to fragment the RNA, then the temperature decreased to 4 °C before being removed to RT. A master mix containing 6:1:1 ratios of cDNA Synthesis Premix : 100mM DTT : StarScript Reverse Transcriptase was made and 4 µl added to each sample. The mixture was then heated in a thermocycler at 25 °C for 5 minutes, 42 °C for 20 minutes to allow the synthesis of cDNA, then the temperature decreased to 37 °C before being removed to RT. Immediately, 1 µl of Finishing Solution was added to each reaction and the plate replaced in the thermocycler and incubated at 37 °C for 10 minutes, 95 °C for 3 minutes, then the temperature decreased to 25 °C, before being removed to RT. A second master mix was made up using 15:1 ratio of Terminal Tagging Premix : DNA Polymerase and 8 µl of the master mix added to each reaction before incubating in a thermocycler at 25 °C for 15 minutes, 95 °C for 3 minutes, to terminally tag the cDNA, then the temperature decreased to 4 °C before being removed to RT. The samples were cleaned using AMPure XP beads (Beckman Coulter) following the ScriptSeq standard protocol. Briefly, to the samples, 45 µl of Ampure XP beads were added and incubated at RT for 5 minutes. The sample was placed on

a magnetic block to hold the Ampure XP beads, and once the supernatant had cleared, the supernatant was removed from the sample. The samples were washed twice with 80% Ethanol, resuspended in 24 µl MBGW and the supernatant containing the DNA moved into a new 96 well PCR plate.

The index PCR master mix was made up in a new plate with 25 µl FailSafe PCR PreMix E, 1 µl Forward PCR Primer, 1 µl Index PCR Primer (each sample was indexed with a different index PCR Primer) and 0.5 µl FailSafe PCR Enzyme per well and 22.5 µl di-tagged cDNA added to each well. The samples were then amplified using the following PCR conditions: 95 °C for 1 minute, followed by 20 cycles of 95 °C for 30 seconds, 55 °C for 30 seconds, 68 °C for 3 minutes, followed by a final anneal at 68 °C for 7 minutes and hold at 12 °C. Post-PCR samples were cleaned using AMPure XP beads, by addition of 30 µl Agencourt AMPure XP beads and elution in 24 µl MBGW. The supernatant containing the purified di-tagged cDNA was then transferred to a new 96 well PCR plate. The supernatant was cleaned using AMPure XP beads (as in Appendix A).

The final samples were quantified using the Qubit® DNA HS Assay Kit (Invitrogen, Thermo Fisher Scientific, Carlsbad, United States) following the manufacturer's instructions (Life Technologies 2015a). Samples were pooled at equimolar concentrations to create a 4 nM pool, and pool quality checked using the Agilent 2200 TapeStation system (Agilent Technologies, Santa Clara, United States) with High Sensitivity D1000 reagents (Agilent Technologies, Santa Clara, United States) following the manufacturer's instructions (Agilent Technologies 2015), to attain a sample peak between 200-1000bp with no small fragments below 200 bp to allow suitable quality for sequencing. The pool was then denatured using NaOH (Illumina 2018a), combined with 5% PhiX, diluted to 10 pM, and run on a single MiSeq flow cell using the V3 reagents kit (Illumina, San Diego, United States) following the manufacturer's instructions (Illumina 2018b).

2.2.1.2 Quality Control

Process blanks, MBGW put through the same processing as samples, were undertaken for these samples, given their own index and run through the sequencer as a sample. An additional indexing blank, MBGW not run through the processing but given its own index at the index PCR stage, were also run through the sequencer as a sample. All were examined using the tapestation and Qubit for quality purposes prior to sequencing. All samples,

including blanks, were examined for read number and those with low quality reads, or low read numbers were filtered out of the analysis.

The PhiX internal standard was spiked into the final pool and run on the sequencer. The PhiX standard was mapped to the PhiX genome on the MiSeq as part of the standard Illumina workflow to allow for the assessment of the quality of the MiSeq run, in addition to metrics on cluster density and read numbers. This was compared to the average statistics on these metrics for the specific run type (amplicon or metagenomics) runs on the MiSeq at Fera to ensure the quality of the run was of the standard usually obtained.

2.2.1.3 Analysis

Reads from the MiSeq were initially trimmed using Sickle v1.33 (Joshi and Fass 2011) to remove sequence of quality less than Q20 (1 in 100 probability of incorrect base call) and length less than 100 base pairs. To identify the number of reads in each sample belonging to the internal spiked positive control (mengo virus), and norovirus, the trimmed data were mapped to the mengo virus complete genome (NCBI DQ294633.1 with a spurious 55 bp region of tandemly-repeating cytosine removed from the start of the sequence) or the norovirus genome (NCBI NC_001959.2) using bwa mem v0.7.10, and alignments analysed using Samtools v0.1.19. This approach is a more sensitive method for detecting sequences from known viruses of interest than non-targeted assignment and therefore provided more accurate results. A paired t-test was performed in RStudio Version 1.0.136 to compare the total number of reads generated and the number of reads mapping to the mengo virus positive control by each enrichment method. R studio was additionally used to calculate the mean percentage of reads obtained after filtering with sickle and to perform a paired t-test.

2.2.2 Limit of Detection and Method Comparison

2.2.2.1 Sample preparation

MS2 Preparation

Freeze-dried *E. coli* (DSMZ strain 5695) and vacuum-dried MS2 phage (DSMZ strain 13767) were purchased from DSMZ (Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures, Brunswick, Germany).

To the freeze-dried *E. coli,* 500 µl NZCYM broth was added and incubated at RT for 20 minutes to rehydrate the cells. Approximately 100 µl of the rehydrated suspension was added to 5 ml NZCYM broth and at 37°C grown for 24 h, to bulk up cells for subsequent MS2

propagation. A further 100 µl of the rehydrated suspension was spread onto NZCYM agar (Table S1B) plate to assess purity of the culture.

To grow the MS2, the bulked-up suspension of *E. coli* was used to create an *E. coli* lawn by mixing 100 µl of the 24 h culture with 4 ml of melted NZCYM soft agar (Table S1C), which was then poured over a NZCYM agar plate to form an overlay and left at RT for 10 minutes to allow to set. The vacuum-dried MS2 was then placed onto the centre of the overlay and 100 µl NZCYM broth added and left for 10 minutes to rehydrate. The plate was incubated for 18 hours at 37°C. After incubation, 5 ml NZCYM broth was added to the plate and placed at RT in a shaking incubator for 4 hours. The broth was then removed from the plate and centrifuged for to remove debris, before the supernatant was filtered through a 45 µm filter.

To bulk-up the phage, 500 µl fresh 24 hour grown *E. coli* was added to 500 µl of the filtered MS2 supernatant and incubated at RT for 30 minutes. 200 µl of the *E. coli* / MS2 mixture was then added to 4 ml of melted NZCYM soft agar, subsequently poured over a NZCYM agar plate to form an overlay and left at RT for 10 minutes to allow to set. This was repeated four times, and the plates were incubated for 24 hours at 37°C. After incubation the soft agar layer was scraped off and placed into 50 ml centrifuge tubes and centrifuged at 3000 *g for 20 minutes. The supernatant was taken from the centrifuged tubes and passed through a 45 µm filter before quantification.

Quantification of the MS2 was done using a Plaque assay (Adams 1959). A dilution series from -1 to -10 was created of the filtered supernatant collected above. 100 µl fresh 24 hour grown *E. coli* was added to 4 ml of melted NZCYM soft agar, poured over a NZCYM agar plate to form an overlay and left at RT for 10 minutes to allow to set. For each MS2 dilution, 10 µl of the phage dilution was spotted on the surface of the plate in triplicate. Three dilutions were done per plate. The plates were incubated for 18 hours at 37°C and then removed and plaques counted for each dilution to calculate the titre of MS2 in Plaque Forming Units (PFU). The MS2 suspension was aliquoted and kept at 4°C for temporary storage (less than 1 week) and -80°C for long term storage (greater than 1 week).

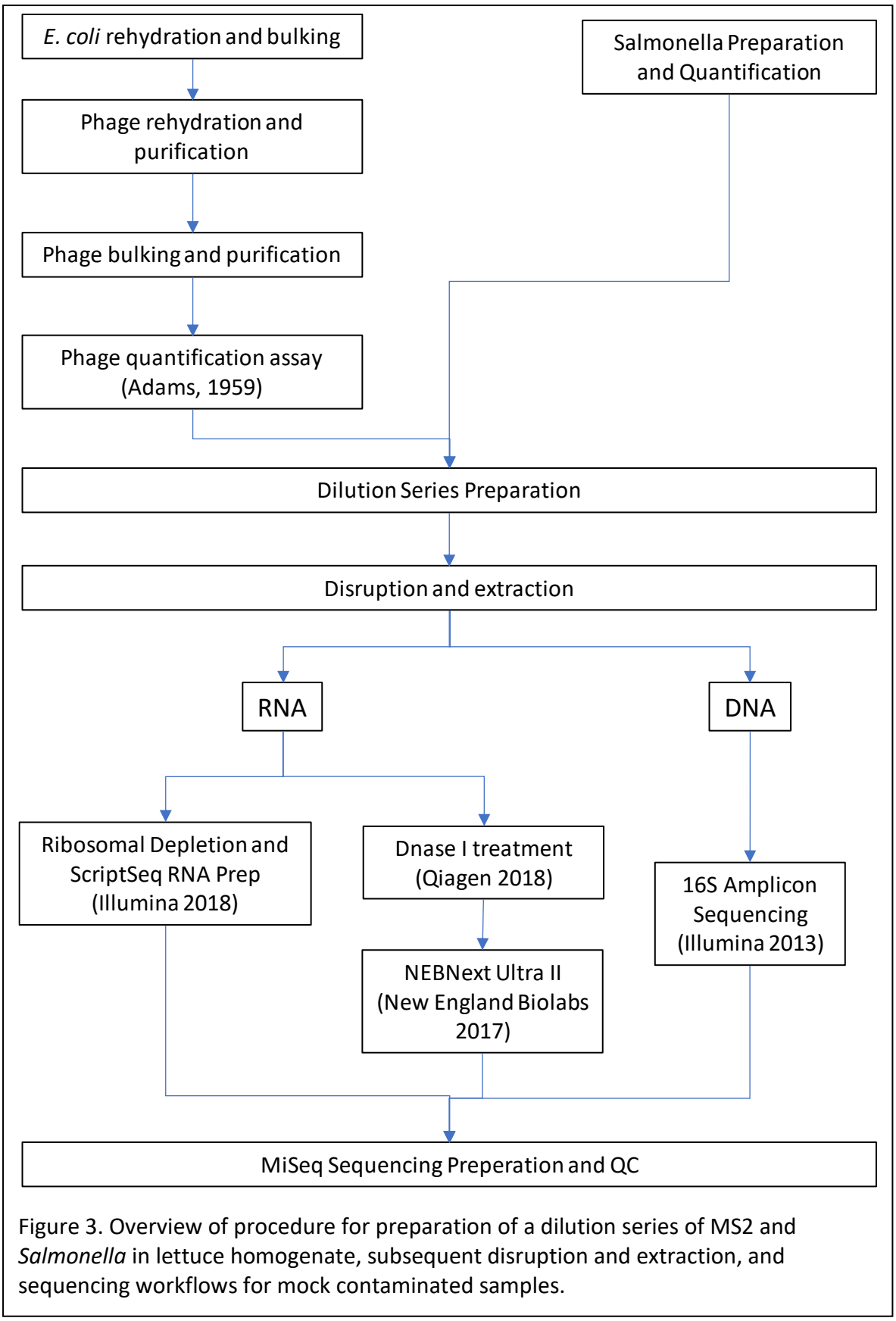


Figure 3. Overview of procedure for preparation of a dilution series of MS2 and *Salmonella* in lettuce homogenate, subsequent disruption and extraction, and sequencing workflows for mock contaminated samples.

Salmonella Preparation

Salmonella enterica subsp. *enterica* serotype Cerro (NCTC 5801) was grown in nutrient broth (NB). A dilution series of the fresh 24-hour culture was created by adding 1 ml of the liquid culture to 9 ml PBS, mixed and then a serial dilution created from -1 to -8. The dilution series from -5 to -8 were plated in duplicate onto nutrient agar (NA), to quantify colony forming units (CFU) per ml and assess purity, through the addition of 100 µl of the appropriate dilution to the plate, spreading using a sterile spreader and incubated at 38 °C for 24 hours. After incubation colonies were counted to allow quantification in colony forming units (CFU). The *Salmonella* culture was kept at 4°C for temporary storage (less than 1 week) and -80°C for long term storage (greater than 1 week).

2.2.2.4 Dilution Series Preparation and Extraction

Iceberg lettuce was purchased from a local retailer and used to make a lettuce homogenate for preparation of a dilution series (Figure 3). Briefly the outermost leaves were removed and placed into a Bioreba filtered grinding bag (Lynchwood Diagnostics, Peterborough, UK) and ground before 24 aliquots each of 50 µl were transferred into sterile 1.5 ml tubes. A serial dilution of MS2 and *Salmonella* was prepared based on the PFU and CFU calculated for the cultures, respectively, through the dilution of the culture in molecular biological grade water (MBGW) and combined with the lettuce homogenate. Each dilution was performed in triplicate. The samples were extracted using the AllPrep DNA/RNA Mini Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions (Qiagen 2005). The RNA was eluted in 50 µl, followed by a further 30 µl of MBGW, and the DNA eluted in 100 µl buffer EB. The eluate was transferred to a new 1.5 ml tube and stored at -80 °C.

2.2.2.3 Sample QC

DNA samples were subject to PCR and real time PCR using the methods outlined in Figure 3. For conventional PCR, a mastermix was made using 12.5 µl, 2* ReddyMix (Thermo Fisher Scientific, Waltham, United States), 5 µl MGW, 1.25 µl 10 mM forward primer Salm-invA-285-F (GTGAAATTATCGCCACGTTCGGGCAA), 1.25 µl 10 mM reverse primer Salm-invA-285-R (TCATCGCACCGTCAAAGGAACC) per sample. This was then run on a thermocycler at 95 °C for 2 mins, followed by 35 cycles of 95 °C for 20 s, 60 °C for 30 s, 72 °C for 90 s, and one cycle at a final annealing temperature of 72 °C for 5 mins. The PCR products were visualised on a 1% agarose gel made with 1 * Tris/Borate/EDTA (TBE) buffer, with 3% ethidium bromide to visualise, and run for 90 mins at 80V, before using UV to image. Real time PCR was

undertaken using the *mericon Salmonella* spp. Kit (Qiagen, Hilden, Germany) following manufacturer's instructions (Qiagen 2012a). Briefly, a mastermix containing 5.4 µl ROX dye and 130 µl Multiplex PCR MM was made and per well 10.4 µl of this mixed with 9.6 µl of either sample, positive control DNA or MGW as required. This was run on a real time PCR instrument, with an initial denaturation of 5 mins at 95 °C, followed by 40 cycles of 95 °C for 15 s, 60 °C for 30 s and 72 °C for 10 s.

2.2.2.4 Sequencing

ScriptSeq RNA-seq

The RNA samples were quantified using the Qubit® RNA HS Assay Kit (Life Technologies 2015b) on the Qubit v2 to ensure an input concentration of lower than 5 µg as specified by the ScriptSeq kit protocol. The samples were then processed for sequencing using the ScriptSeq Compete Plant Kit (Illumina, San Diego, United States) following the manufacturer's protocol (Illumina 2018d), as outlined in section 2.2.1.1 workstream i. Briefly, this included treatment with Ribozero to remove ribosomal RNA, synthesis and tagging of cDNA and addition of unique indexes by PCR.

Post-PCR samples were cleaned using Agencourt AMPure XP beads (Beckman Coulter, Brea, United States) following the protocol presented in Appendix A, with the addition of 30 µl of Ampure XP beads and elution in 35 µl MBGW. The samples were then quantified using the Qubit® DNA HS Assay Kit (Life Technologies 2015a) and pooled at equimolar concentrations to create a 4 nM pool. Pool quality was checked using the Agilent 2200 TapeStation system with High Sensitivity D1000 reagents (Agilent Technologies 2015). The sample peak was between 200 bp and 1000 bp, with no small fragments below 200 bp, thus meeting the required criteria for MiSeq analysis. The 24 sample 4 nM pool was denatured and combined with PhiX (Illumina 2018a) then run on the MiSeq using the V3 reagents kit at 8 pM with 5% PhiX (Illumina 2018b).

NEB-Next Ultra II RNA-seq

The RNA samples were subject to treatment with DNase I (Qiagen, Hilden, Germany) to minimise DNA contamination prior to sequencing. This required mixing 30 µl of sample with 13.75 µl MBGW and addition of 5 µl Buffer RDD, and 1.25 µl DNase I. The mix was incubated at room temperature for 10 mins and then cleaned-up using AMPure XP beads (Appendix A) with an input of 90 µl AMPure XP beads (ratio 1.8:1, beads to sample), and elution in 20 µl of MBGW.

The DNA-free RNA was then prepared for sequencing using the NEBNext® Ultra™ II RNA Library Prep Kit for Illumina® following the manufacturer's protocol NEBNext® Ultra™ II RNA Library Prep Kit for Illumina Instruction manual, version 1, chapter 2 (New England Biolabs 2017). To 12 µl RNA, 1 µl rRNA depletion solution and 2 µl buffer were added and mixed before placing on a PCR machine for 2 min at 95 °C, ramp down 0.1 °C per s to 22 °C, held 22 °C for 5 min. After the hold, 2 µl RNAse H, 2 µl Buffer and 1 µl MBGW were added and the mix incubated at 37 °C for 30 min. To this, 2.5 µl DNAse I, 5 µl buffer, 22.5 µl MBGW were added to remove DNA and incubated at 37 for 30 min. To remove DNAse and clean AMPure XP beads were used as per Appendix A, eluting in 7 µl MBGW. To enrich the RNA, 4 µl buffer and 1 µl random primers were added to the eluate and incubated for 8 min at 94 ^{o}C. After incubation, 8 µl MBGW and 2 µl First strand synthesis enzyme mix were added, and then the mix thermal cycled for 10 min at 25, 50 min at 42 ^{o}C, 15 min at 70 ^{o}C and held at 4 ^{o}C. A further 8 µl buffer, 4 µl Second strand synthesis enzyme and 48 µl MBGW were added and the mix incubate for 1 hour at 16 ^{o}C. The product was cleaned using AMPure XP beads (as in Appendix A), eluting in 50 µl 0.1x TE buffer. To this, 7 µl buffer, and 3 µl End Prep enzyme mix were added and incubated for 30 min at 20 and 30 min at 65 ^{o}C. A 5-fold dilution of NEBNext adapter in buffer was made and 2.5 µl of this dilution was combined with 1 µl ligation enhancer and 30 µl ligation master mix before being added to the sample and incubated for 15 min at 20 ^{o}C. To this, 3 µl USER enzyme was added and incubate for a furhher 15 min at 37 ^{o}C. The product was cleaned using AMPure XP beads, eluting in 15 µl 0.1x TE buffer, before being combined with 25 µl Q5 master mix and 5 µl of each of forward and reverse primers and thermocycled for 30s at 98 ^{o}C, then 10 cycles of 98 ^{o}C for 10 s, 65 ^{o}C for 75 s, followed by a final 5 min at 65 ^{o}C. The final samples with adapters were cleaned using AMPure XP beads, eluting into 20 µl 0.1x TE buffer.

The 24 samples were quantified using the Qubit® DNA HS Assay Kit (Life Technologies 2015a) and pooled to equimolar concentrations. Pool quality was checked using the Agilent 2200 TapeStation system with High Sensitivity D1000 reagents (Agilent Technologies 2015) and quantified using the Qubit® DNA HS Assay Kit following the manufacturer's instructions. The 24 sample 4 nM pool was denatured and combined with 5% PhiX (Illumina 2018a) and was run on the MiSeq using the V3 reagents kit at 8 pM (Illumina 2018b).

16S rRNA Gene Amplicon Sequencing

The DNA samples underwent preparation for 16S rRNA gene amplicon sequencing (see Figure 3). PCR reactions of 30 µl were carried out using the Phusion High-Fidelity DNA Polymerase (New England Biolabs) containing 6 µl of HF buffer, 0.3 µM forward (TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTGYCAGCMGCC-GCGGTAA) and reverse (GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGACTACNVGGGTWTCTA-AT) primers (Caporaso *et al.* 2011) containing Nextera tags, 0.3 mM of dNTPs, 0.3 µl Phusion DNA polymerase and 1 µl of 1:10 diluted template DNA. The final reaction volume was made up with nuclease-free water. A touchdown PCR protocol was followed. Amplification started with an initial single denaturation step for 2 min at 98°C, followed by 22 cycles of denaturation at 98°C for 20 s, annealing for 45 s starting at 65°C with a reduction of 0.5°C per cycle down to 54°C, and extension for 60 s at 72°C. This was followed by a further 8 cycles of denaturation for 20 s at 98°C, annealing for 45 s at 54°C, and extension for 60 s at 72°C prior to a final extension for 10 min at 72°C. PCR Products were cleaned-up using AMPure XP Beads (Appendix A) with an input of 24 µl of beads, and elution in 50 µl MBGW. Nextera XT sequencing adapters and indexes (Illumina) were then attached using Phusion High-Fidelity DNA Polymerase by combining 10 µl HFb, x 0.3 µM dNTP, x 1 µM $MgCl_2$, 0.5 µl Phusion polymerase and 5 µl of each unique index 1, 2 and the purified PCR product per sample. The final reaction volume was made up with nuclease-free water. Amplification started with an initial denaturing step for 3 mins at 95°C, followed by 8 cycles of denaturation for 30 s at 95°C, annealing for 30 s at 55°C and extension for 30 s at 72°C prior to a final extension for 5 mins at 72°C. The PCR product was then cleaned-up using AMPure XP Beads (Appendix A), with addition of 56 µl of beads and elution in 25 µl MBGW. All samples were then quantified using the Qubit® DNA HS Assay Kit (Life Technologies 2015a) and pooled to equimolar concentrations with a further 96 unrelated amplicon samples to reflect a standard amplicon run. Pool quality was checked using the Agilent 2200 TapeStation system with High Sensitivity D1000 reagents (Agilent Technologies 2015). The 120-sample pool was denatured with NaOH, plus an additional heat denaturation step, then combined with 10% PhiX (Illumina 2018a) and run on the MiSeq using the V3 reagents kit at 10 pM (Illumina 2018b).

2.2.2.5 Quality Control

Extraction blanks were undertaken as part of the RNA extraction. Process blanks, MBGW put through the same processing as samples, were undertaken for each sequencing method. All were examined using the tapestation and Qubit for quality purposes. For ScriptSeq and 16S rRNA sequencing these were run on the sequencer as a separate sample. For NEB, due to a lack of availability of unique indexes, the negatives were unable to be sequenced. An additional indexing blank, MBGW not run through the processing but given its own index at the index PCR stage, were also done for the 16S rRNA method and run through the sequencer as a separately indexed sample. All samples, including blanks, were examined for read number and those with low quality reads, or less than 500 reads were filtered out of the analysis. Any blanks remaining were run through the bioinformatic analysis separately to samples and the top taxa compared manually to those in experimental samples to rule out cross contamination.

The PhiX internal standard was spiked into the final pool and run on the sequencer. The PhiX standard was mapped to the PhiX genome on the MiSeq as part of the standard Illumina workflow to allow for the assessment of the quality of the MiSeq run, in addition to metrics on cluster density and read numbers. This was compared to the average statistics on these metrics for the specific run type (amplicon or metagenomics) runs on the MiSeq at Fera to ensure the quality of the run was of the standard usually obtained.

2.2.2.6 Bioinformatics

RNA-Seq Analysis

Several methodologies were undertaken to interrogate RNA-Seq data to examine the impacts of bioinformatic protocols on LoD. Initially, reads from the MiSeq were trimmed using Sickle v1.33 (Joshi and Fass 2011) to remove sequence of quality less than Q20 (1 in 100 probability of incorrect base call) and lengths less than 100 base pairs. The remaining reads were used in subsequent analyses using: MG-RAST v.4.0.3 (Meyer *et al.* 2008), Kraken v1 (Wood and Salzberg 2014) with and without an initial prefiltering step to remove lettuce contamination, Bracken (Lu *et al.* 2017) with and without an initial prefiltering step to remove lettuce sequence contamination and mapping to the *Salmonella* and MS2 genomes using BWA-MEM (Li and Durbin 2009).

Kraken was also used to assign taxonomy following the mapping of reads to the lettuce chloroplast. Mapping was done using BWA-MEM v0.7.10 and then data filtered using

Samtools v0.1.19 (Li *et al.* 2009) to give a list of the unmapped reads. The unmapped reads were assigned taxonomy using MiniKraken and full Kraken databases. The outputs were summarised in Kraken-summary tables and used to generate summary tables of read count against assignment. Summary data from Kraken outputs was then used in Bracken v1.0.0 to create a semi-quantitative list of the species present in each sample.

Trimmed data were mapped to the *Salmonella* and MS2 complete genomes using BWA-MEM, and alignments were analysed using Samtools to create an output of the number of reads mapping to each of the genomes.

16S rRNA gene Sequence Analysis

The data were analysed using several methods: MG-RAST, QIIME (Caporaso *et al.* 2010), mapping to the *Salmonella* genome using BWA-MEM and QIIME2 (Bolyen *et al.* 2019). MG-RAST 16S rRNA gene sequencing samples were QC'd using the MG-RAST standard online pipeline and taxonomy assigned using the Greengenes database (DeSantis *et al.* 2006). Analysis by QIIME v1.9.0 briefly comprised QC to remove reads of low quality or length, the removal of chimeric sequences from the data, OTU picking and taxonomic assignment, then visualisation within QIIME. Quality controlled data were mapped to the *Salmonella* complete genome using BWA-MEM, and alignments analysed using Samtools to create an output of the number of reads mapping to the genome. QIIME2 was adopted using methodology outlined in the tutorial documentation, all steps outlined were performed as part of the QIIME2 pipeline. Importing of data was done following the Cassava 1.8 paired-end demultiplexed fastq section of the importing tutorial (Qiime2docs 2017a). The "Moving Pictures" tutorial (Qiime2docs 2017b) was then followed from the summary of demultiplexed results onwards. DADA was used to denoise the data then sequences were aligned using MAFFT (Nakamura *et al.* 2018) and filtered to remove variable positionings. A tree was generated, rooted and used to analyse diversity. Diversity was analysed using alpha rarefaction and beta significance. Taxonomic assignment was completed using the command "qiime feature-classifier classify-sklearn".

Assessment of LoD

For all methodologies, species with fewer than 10 reads were filtered out of the data and are reported as having zero reads for subsequent analysis. The LoD was assigned as the lowest concentration at which greater than 10 reads were observed for the target species for all replicates at that concentration. Basic statistics to examine the linearity of the relationship

between the number of reads assigned and the concentration of MS2 or *Salmonella* spiked were performed in Microsoft Excel (Office 365, Microsoft, Redmond, United States).

2.2.3 MiSeq vs HiSeq read simulation

To create a simulated mock community, the top bacterial genera from the analysis of the LoD lettuce microbiome samples were downloaded from the NCBI database (Table 4) and combined with the lettuce chloroplast genome (NC 007578.1 *Lactuca sativa* chloroplast, complete genome) to simulate host contamination. Varying levels of *Salmonella enterica* (NC 003198.1 *Salmonella enterica* subsp. enterica serovar Typhi str. CT18, complete genome) were also added to the mock community to create a dilution series, giving a final abundance of *Salmonella* ranging from 0 to 10% of the sample (Appendix B). This was then used to simulate 1 million MiSeq, HiSeq or NovaSeq reads for each "dilution", performed in triplicate, using InSilicoSeq (Gourle *et al.* 2018). The 21 mock samples were then analysed using Kraken with the mini database. Read numbers were assessed in Microsoft Excel and a single factor ANOVA performed to check the significance of the difference in read numbers between platforms.

Table 4. NCBI accession number, strain name, and details on chromosome or full genome used, for all isolates used to produce the mock community dilution series.

Accession Number	Name	Genome detail
NC 002944.2	*Mycobacterium avium* subsp. paratuberculosis str. k10	complete genome
NC 002947.4	*Pseudomonas putida* KT2440 chromosome	complete genome
NC 004722.1	*Bacillus cereus* ATCC 14579 chromosome	complete genome
NC 016830.1	*Pseudomonas fluorescens* F113	complete genome
NZ CP010519.1	*Streptomyces albus* strain DSM 41398	complete genome
NZ CP011007.1	*Bacillus pumilus* strain SH-B9	complete genome
NZ LT700188.1	*Negativicoccus massiliensis* strain Marseille-P2082 genome assembly	chromosome: contig00001
NC 003198.1	*Salmonella enterica* subsp. enterica serovar Typhi str. CT18	complete genome

2.3 Results

2.3.1 Quality Assessment

Run quality metrics are found in Table 5. All runs had a high level of reads passing filter and assigned to indexes, indicative of a good quality run. Negative controls were examined and all had read numbers of less than 500, therefore were filtered out at QC stage and no subsequent analysis was performed on them.

Table 5. MiSeq run metrics for each run associated with data from enrichment comparison and limit of detection (LoD) study

Method	Cluster Density	Reads Passing Filter	% of Clusters Passing Filter	% PhiX Loaded into Library	Concentration of library loaded	% of Reads Aligned to PhiX
Enrichment Comparison	1168	24,185,876	89.13	5	8	7.17
LoD: ScriptSeq	1021	21,308,336	89.66	5	10	4.33
LoD: NEBNext	1166	25,370,000	91.37	5	10	2.91
LoD: 16S	866	18,635,296	83	10	10	4.61

Method	Error Rate	%Q30	% Identified	% Assigned to Index	Number of Samples in Pool
Enrichment Comparison	3.19	46.36	92.9	99	27
LoD: ScriptSeq	3.35	69.05	92.51	96	24
LoD: NEBNext	4.52	63.10	97.44	100	24
LoD: 16S	3.99	63.43	81.9	86	120

2.3.2 Enrichment Comparison

All samples contained less RNA than the minimum input recommended for the Scriptseq complete protocol. Nevertheless, through use of the low input adaptations to the protocol and increased cycle numbers, a total of 44,898,822 reads were produced, with data meeting QC requirements obtained for all samples (Table 6). The number of reads produced was significantly ($p = 0.004$, paired t-test) less for poly-A capture than for ribosomal depletion. The percentage of reads retained following QC using Sickle was also significantly ($p \leq 0.0001$, paired t-test) different between the two enrichment methodologies with the mean value for retained sequence for poly-A capture and for ribosomal depletion being 5% and 39%

respectively. This is indicative of the input RNA quality being low or fragmented or of methodological incompatibilities between the enrichment and sequencing methods.

No reads mapped to the norovirus genome for either enrichment methodology, and the number of reads mapping to mengo virus (see Table 6) was significantly ($p= 0.02$, paired t-test) different between enrichment methods. This finding indicates that ribosomal depletion may provide a more sensitive means of detection than poly-A capture when targeting foodborne viruses.

Table 6. Total number of reads and those passing quality and length filters, plus number of reads mapping to mengo virus for each sample using the poly-A capture and ribosomal depletion enrichment methodologies.

Sample	Enrichment method	Total number of reads	Total number of reads post-trim	Reads mapped to mengo virus
1	Poly-A capture	1199920	42830	0
1	Ribosomal depletion	1571760	538156	0
2	Poly-A capture	1115124	74568	0
2	Ribosomal depletion	1095900	233798	4
3	Poly-A capture	1006296	43914	0
3	Ribosomal depletion	1473212	404342	2
4	Poly-A capture	995802	49236	0
4	Ribosomal depletion	2139938	750424	4
5	Poly-A capture	956490	69938	0
5	Ribosomal depletion	3051470	1351222	2
6	Poly-A capture	1218258	59384	0
6	Ribosomal depletion	2966918	1049910	4
7	Poly-A capture	835690	46934	4
7	Ribosomal depletion	5857586	3747664	0
8	Poly-A capture	686446	45954	0
8	Ribosomal depletion	3975074	1869944	2
9	Poly-A capture	1260962	37590	0
9	Ribosomal depletion	1553518	512870	2
10	Poly-A capture	613204	24176	0
10	Ribosomal depletion	3610274	1510350	0
11	Poly-A capture	822908	45916	0
11	Ribosomal depletion	2802394	1241134	2
12	Poly-A capture	1221984	37818	0
12	Ribosomal depletion	1033246	351834	0

2.3.3 Limit of Detection Methods Comparison

2.3.3.1 Sample QC results

Using conventional PCR with visualisation on an agarose gel, the lowest detectable level of *Salmonella* in the samples was 10^3 CFU/ml (Figure 4). The qPCR method tested proved more sensitive, detecting *Salmonella* consistently at levels down to 10^2 CFU/ml (Figure 5).

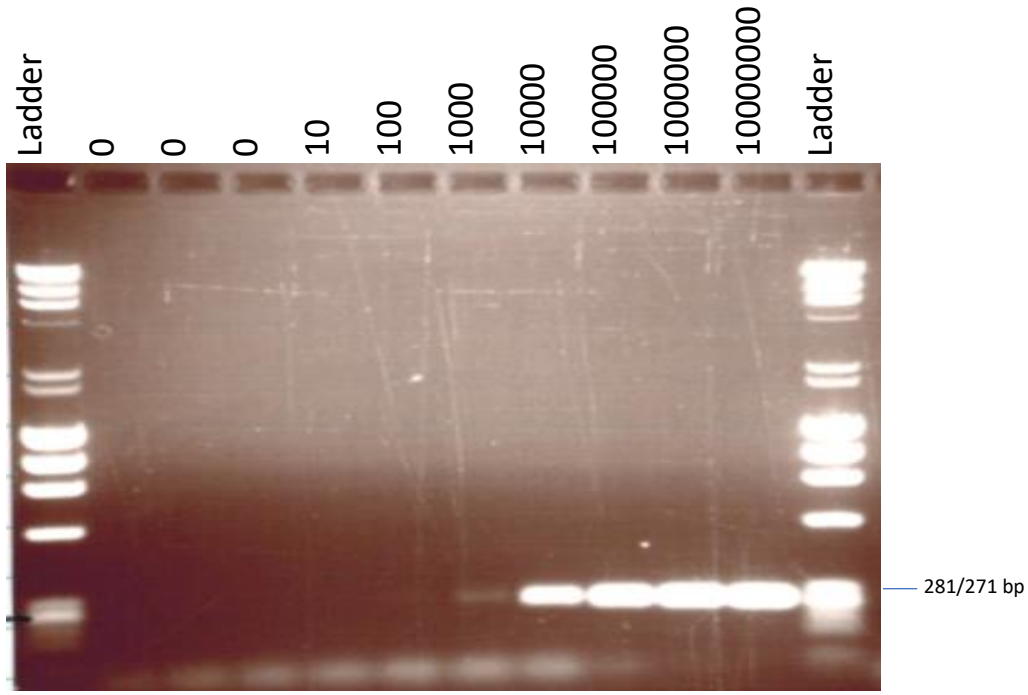


Figure 4. Agarose gel image showing PCR products for *Salmonella* specific PCR on limit of detection dilution series DNA extracts. Samples are labelled with concentration of *Salmonella*.

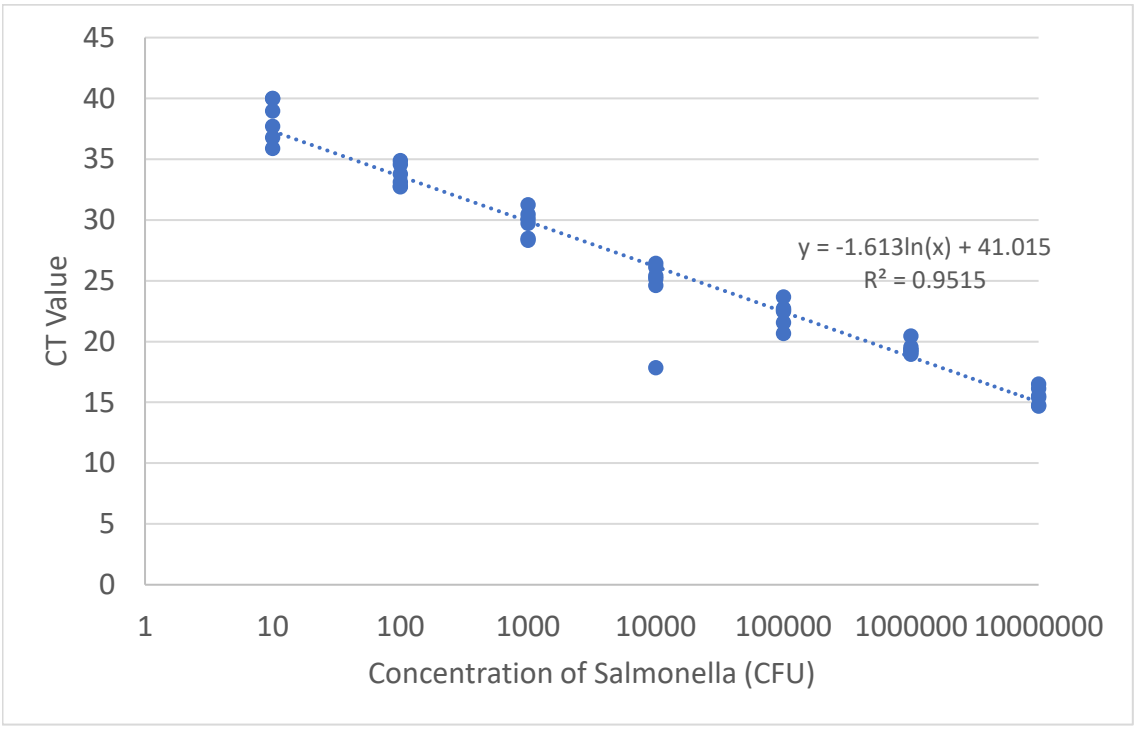


Figure 5. *Salmonella* qPCR results (CT) for dilution series samples. Samples not spiked with *Salmonella* had CT values that were undetermined.

2.3.3.2 Bioinformatics Method Comparison

The number of reads of *Salmonella* and MS2 detected using the different sequencing and bioinformatic approaches are summarised in Figure 6 with the linearity of relationships examined in Table 7. Table 8 shows the relationship between the methods and concentration of pathogen in samples and the mis-assignment of *Salmonella*, as well as outlining the LoDs for each approach.

A: MS2 using ScriptSeq

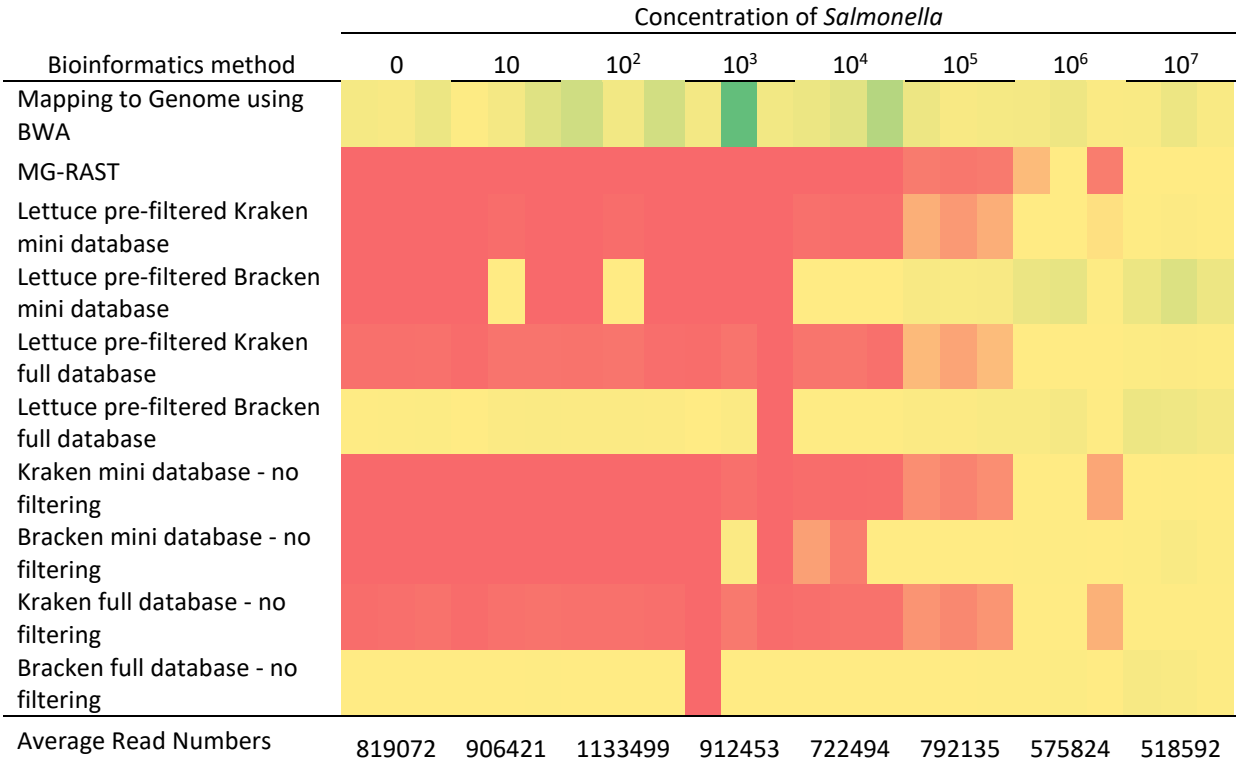
Bioinformatics method	Concentration of MS2							
	0	10	10^2	10^3	10^4	10^5	10^6	10^7
Mapping to Genome using BWA								
MG-RAST								
Lettuce pre-filtered Kraken mini database								
Lettuce pre-filtered Bracken mini database								
Lettuce pre-filtered Kraken full database								
Lettuce pre-filtered Bracken full database								
Kraken mini database - no filtering								
Bracken mini database - no filtering								
Kraken full database - no filtering								
Bracken full database - no filtering								
Average Read Numbers	819072	906421	1133499	912453	722494	792135	575824	518592

B: MS2 using NEBNext

Bioinformatics method	Concentration of MS2							
	0	10	10^2	10^3	10^4	10^5	10^6	10^7
Mapping to Genome using BWA								
MG-RAST								
Lettuce pre-filtered Kraken mini database								
Lettuce pre-filtered Bracken mini database								
Lettuce pre-filtered Kraken full database								
Lettuce pre-filtered Bracken full database								
Kraken mini database - no filtering								
Bracken mini database - no filtering								
Kraken full database - no filtering								
Bracken full database - no filtering								
Average Read Numbers	937986	1070995	1177812	1244793	647349	1231444	1078245	851226

Figure 6 (continued on next page)

C: *Salmonella* using ScriptSeq



D: *Salmonella* using NEBNext

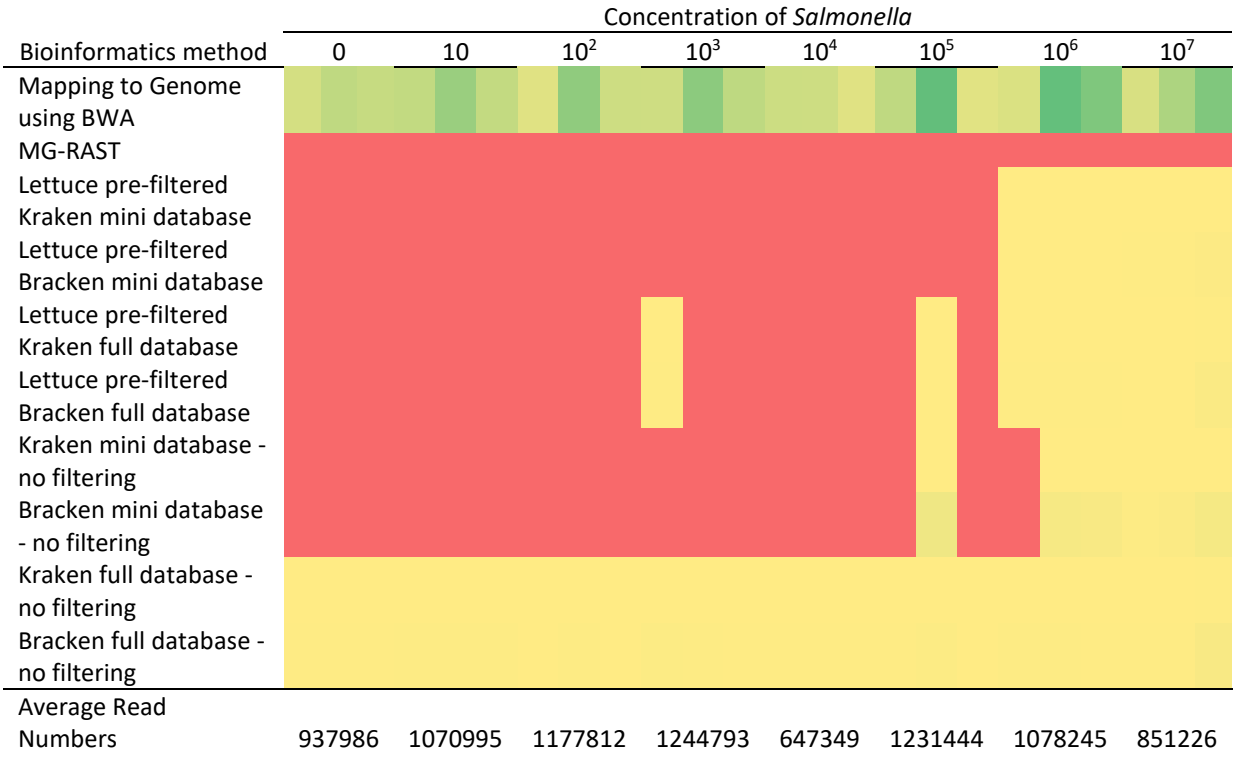


Figure 6 (continued on next page)

E: 16S

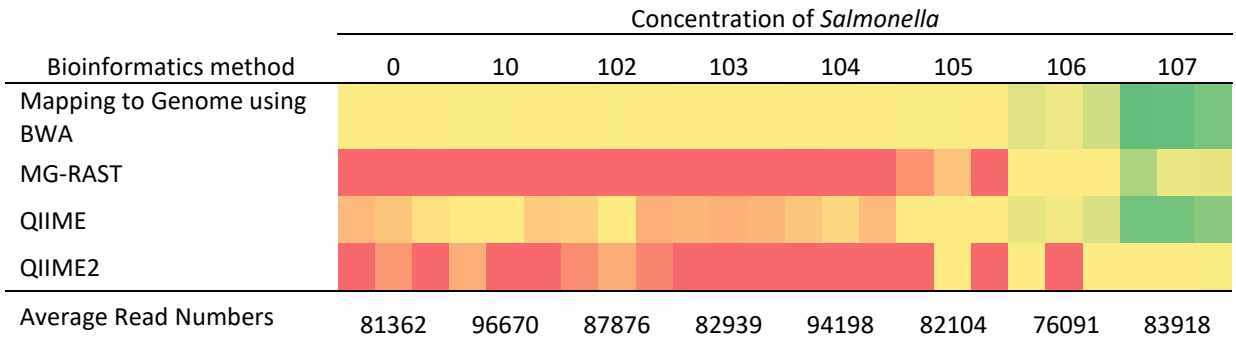


Figure 6. Heat map showing number of reads for different bioinformatics procedures for A: MS2 using the ScriptSeq methodology B: MS2 using the NEBNext methodology C: *Salmonella* using the ScriptSeq methodology D: *Salmonella* using the NEBNext methodology E: *Salmonella* using the 16S rRNA gene sequencing methodology. Red indicates a read number less than 10; yellow a read number between 10 and 500; and green a read number greater than 500.

Table 7. Linear regression parameters for concentration of target against read number; where y-formula = ax + b, and shows the slope and y-coordinates, and R2 = regression coefficient. Analysis method A: RNAseq data, B: 16S rRNA gene amplicon data.

A

	ScriptSeq MS2		ScriptSeq *Salmonella*		NEBNext MS2		NEBNext *Salmonella*	
Analysis Method	y formula	R^2	y formula	R^2	y formula	R^2	y formula	R^2
Mapping to Genome using BWA	0.0002x + 6.6018	0.799	-0.0054x + 79161	0.0376†	1E-05x - 0.1849	0.9237	0.004x + 189698	0.0164†
MG-RAST	5E-05x + 3.6701	0.8299	0.0002x + 15.22	0.8191	0*	na	0	na
Lettuce pre-filtered Kraken mini database	0.0002x + 5.9136	0.7916	0.0007x + 75.628	0.839	1E-05x - 0.1554	0.9213	2E-05x + 1.2138	0.8523
Lettuce pre-filtered Bracken mini database	0.0002x + 5.9136	0.7916	0.0065x + 8425.9	0.5784	1E-05x - 0.1554	0.9213	0.0006x - 4.196	0.7983
Lettuce pre-filtered Kraken full database	0.0002x + 6.2109	0.7976	0.0009x + 82.933	0.9399	1E-05x - 0.1849	0.9237	0.0002x + 102.96	0.3008
Lettuce pre-filtered Bracken full database	0.0002x + 6.2109	0.7976	0.0036x + 9873.4	0.7532	1E-05x - 0.1849	0.9237	0.0005x + 600.92	0.3935
Kraken mini database - no filtering	9E-05x + 2.9014	0.7919	0.0004x + 40.607	0.8367	6E-06x - 0.4377	0.9181	1E-05x + 1.0583	0.833
Bracken mini database - no filtering	9E-05x + 2.9014	0.7919	0.0011x + 1188.4	0.5389	6E-06x - 0.4377	0.9181	0.0012x + 3770.3	0.1322
Kraken full database - no filtering	9E-05x + 3.19	0.7966	0.0005x + 48.808	0.9405	6E-06x - 0.4426	0.9214	0.0002x + 163.09	0.2677
Bracken full database - no filtering	9E-05x + 3.19	0.7966	0.0015x + 2273.1	0.7139	6E-06x - 0.4426	0.9214	0.0005x + 4750.6	0.196

* indicate values where there are zero regression/na R2 values due no data being obtained for this sample analysis

† indicates non-significant regression at p=0.05

B

	16S rRNA gene amplicon	
Analysis Method	y formula	R^2
Mapping to Genome using BWA	0.0009x + 422.09	0.9763
MG-RAST	0.0003x - 19.451	0.6394
QIIME	0.0009x + 174.63	0.9779
QIIME2	3E-05x + 24.85	0.7875

Table 8. Mean number of reads of *Salmonella* mis-assigned in the lettuce control samples and the limit of detection for *Salmonella* and MS2 for sequencing and bioinformatics methods.

Sequencing Method	Bioinformatics Method	Mean Reads of *Salmonella* in Lettuce Control	Limit of Detection *Salmonella* (CFU extraction^{-1})	Limit of Detection MS2 (CFU extraction^{-1})
ScriptSeq	Mapping to Genome using BWA	35977	Undetectable*	10^5
ScriptSeq	MG-RAST	0	10^5	10^5
ScriptSeq	Lettuce pre-filtered Kraken mini database	0	10^4	10^5
ScriptSeq	Lettuce pre-filtered Bracken mini database	0	10^4	10^5
ScriptSeq	Lettuce pre-filtered Kraken full database	26	Undetectable*	10^5
ScriptSeq	Lettuce pre-filtered Bracken full database	8069	Undetectable*	10^5
ScriptSeq	Kraken mini database - no filtering	0	10^4	10^5
ScriptSeq	Bracken mini database - no filtering	0	10^4	10^5
ScriptSeq	Kraken full database - no filtering	23	Undetectable*	10^5
ScriptSeq	Bracken full database - no filtering	2055	Undetectable*	10^5
NEBNext	Mapping to Genome using BWA	146104	Undetectable*	10^7
NEBNext	MG-RAST	0	Undetectable	Undetectable
NEBNext	Lettuce pre-filtered Kraken mini database	0	10^6	10^7
NEBNext	Lettuce pre-filtered Bracken mini database	0	10^6	10^7
NEBNext	Lettuce pre-filtered Kraken full database	0	10^6	10^7
NEBNext	Lettuce pre-filtered Bracken full database	0	10^6	10^7
NEBNext	Kraken mini database - no filtering	0	10^7	10^7
NEBNext	Bracken mini database - no filtering	0	10^7	10^7
NEBNext	Kraken full database - no filtering	50	Undetectable*	10^7
NEBNext	Bracken full database - no filtering	3597	Undetectable*	10^7
16S	Mapping to Genome using BWA	299	Undetectable*	Undetectable
16S	MG-RAST	0	10^6	Undetectable
16S	QIIME	56	Undetectable*	Undetectable
16S	QIIME2	9	Undetectable*	Undetectable

* LoD was unable to be determined due to incorrect assignment of *Salmonella* in lettuce control samples

For the detection of MS2 phage, all bioinformatics methods tested yielded the same LoD; 10^5 PFU/reaction for the ScriptSeq sequencing and 10^7 PFU/reaction for the NEBNext sequencing (Figure 6). No false positive results were detected using any method. Detection of MS2 using ScriptSeq and NEBNext both yielded a linear relationship between the concentration of MS2 and the number of reads assigned to MS2 for all bioinformatic methods tested (Table 7), except for NEBNext in conjunction with MG-RAST which was unable to detect any MS2 within the samples. At the highest concentration the number of reads of MS2 detected varied considerably (Figure 6). For ScriptSeq, the highest average for the three replicates was 1860 reads at 10^7 PFU/reaction, by mapping to the reference using BWA, and the lowest average for the three replicates was 486 reads at 10^7 PFU/reaction, using MG-RAST. For NEBNext, the highest average read number for the three replicates was 119 reads at 10^7 PFU/reaction, found using multiple methods: mapping to the reference using BWA and both Kraken and Bracken in combination with prefiltering of lettuce (Figure 6).

For the detection of *Salmonella,* using both the ScriptSeq and NEBNext sequencing protocols there were differences in LoD dependent upon bioinformatics techniques utilised, with the most sensitive technique delivering an LoD of 10^4 CFU/reaction. Many of the bioinformatics techniques used delivered incorrectly assigned reads of *Salmonella* in the lettuce homogenate unspiked with *Salmonella* and in which no *Salmonella* had been detected during QC (Section 3.2.1). Only Kraken and Bracken in conjunction with the mini database, and MG-RAST, yielded no mis-assigned *Salmonella* reads. These methods also produced a linear relationship between read number and the quantity of *Salmonella* spiked into the samples. For the detection of *Salmonella* using 16S rRNA gene sequencing, all methods tested, bar assignment using MG-RAST, yielded incorrect assignment of *Salmonella* in the unspiked samples. MG-RAST delivered an LoD of 10^6 CFU/reaction. When the mis-assigned reads were subtracted from the number of reads found in the samples, QIIME gave a more sensitive LoD of 10^5 CFU/reaction, QIIME2 gave a less sensitive LoD of 10^7 CFU/reaction, and BWA gave an LoD of 10^6 CFU/reaction. All methods showed a linear relationship between the number of reads and the CFU of *Salmonella* in the original sample, including mapping to the reference genome using BWA which contrasts with results for ScriptSeq and NEBNext.

Overall, the bioinformatics method yielding the most sensitive LoD was prefiltration to remove lettuce associated reads, followed by the assignment of species using Kraken,

followed by Bracken. Differences in the total number of sequence reads for each sample accounted for much of the variability in the number of reads associated with MS2 and *Salmonella* between extraction replicates.

2.3.3.3 Sequencing Method Comparison

ScriptSeq sequencing gave the most sensitive LoD for detection of both MS2 (10^5 PFU) and *Salmonella* (10^4 CFU). Although a comparable methodology, the NEBNext sequencing protocol gave a less sensitive LoD, detecting MS2 10^7 PFU at and *Salmonella* at 10^6 CFU. Amplicon sequencing using 16S rRNA gene sequencing is (by its semi-targeted nature) unable to detect MS2, or any non-bacterial species, due to the lack of the 16S rRNA gene. The LoD of *Salmonella* using 16S rRNA gene sequencing was 10^6 CFUs; less sensitive than that of ScriptSeq kit but equivalent to the NEBNext.

2.3.4 Read Simulation Comparison

Despite the differences in error profiles and read length, for the same number of reads, the HiSeq, MiSeq and NovaSeq all yielded the same limit of detection (Figure 7). The number of reads associated with *Salmonella* varied, with the average of the 3 replicates at the highest "concentration" tested being 28132 for the HiSeq, 49207 for the MiSeq and 48554 for the NovaSeq. There was no significant ($p \leq 0.9$, one-way ANOVA) difference in the number of reads between the three platforms. A linear relationship was found between "concentration" of *Salmonella* and the number of reads assigned to *Salmonella* for all platforms. Despite having an input of the equivalent of 100000 reads of *Salmonella* at the highest concentration, using this analysis none of the platforms were able to detect greater than 50% of the reads as belonging to *Salmonella* (Figure 7).

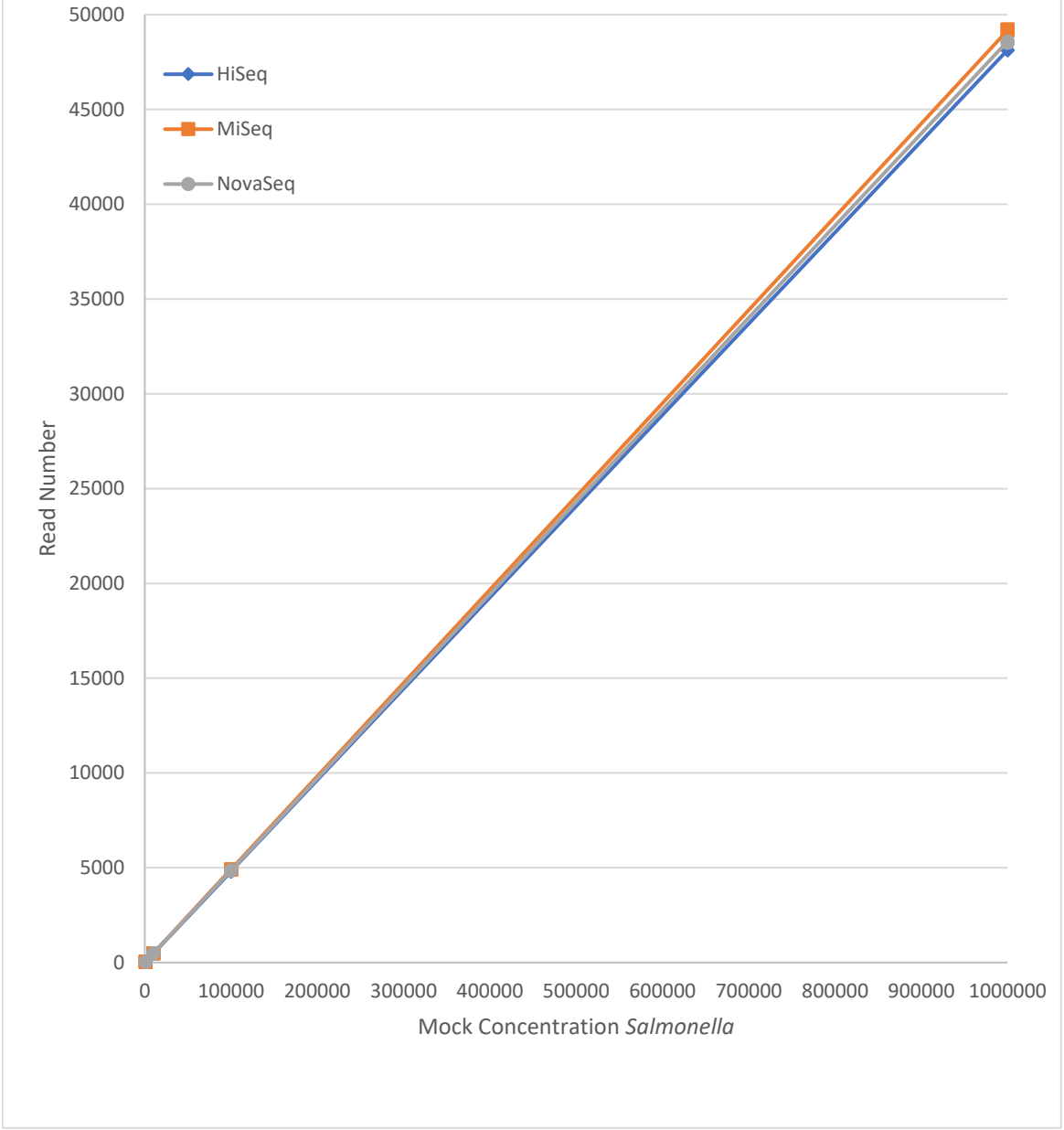


Figure 7. Average number of simulated read counts assigned to *Salmonella* at a mock 'concentration' (equivalent to read number) for three platforms; HiSeq, MiSeq and NovaSeq; that simulated microbiome reads were created for and analysed using Sickle and Kraken

2.4 Discussion

Next Generation Sequencing is increasingly being applied in broader settings, with recent approaches using both amplicon sequencing and metatranscriptomics to study microbial, particularly bacterial, communities associated with fresh produce (Adams *et al.* 2009; Jackson *et al.* 2013; Aw *et al.* 2016; Beckers *et al.* 2016).

Initial results exploring the most appropriate enrichment method to study the microbiome associated with fresh leafy produce showed that ribosomal depletion gave a greater number of reads associated with the control than polyA capture, suggesting it is a superior method for the detection of the microbiome in metatranscriptomics studies. Although sequencing was successfully performed on all samples no norovirus was detected for any qPCR-positive samples, this is likely due to the low titre of the virus and the low RNA input levels. The small quantity of RNA present in the samples prior to processing and the indication of low-quality RNA shown by the filtering out of a large proportion of reads suggests that, although preferable for real time-PCR, the ISO extraction methodology and storage of samples was not appropriate for metagenomics. Therefore, a different extraction method was used for subsequent experiments which employed homogenisation to facilitate harvest of the total microbiome, inclusive of microbes strongly adhered to the surface, associated with biofilms or internalised. This allowed for a non-targeted analysis, unachievable using other techniques, such as washing and recovering rinsates. In addition, the use of a non-specialised column-based kit facilitated the recovery of total RNA and DNA (Leonard *et al.* 2015; Aw *et al.* 2016; Beckers *et al.* 2016).

Work examining the limit of detection of multiple sequencing methods on the MiSeq platform identified that the most sensitive approach delivered an LoD of 10^4 CFU for *Salmonella* and 10^5 PFU for MS2. This approach used the ScriptSeq sequencing method in combination with ribosomal depletion, followed by bioinformatic analysis using Kraken with the mini database. The LoD identified in this study is less sensitive than current methodologies, and cannot detect human pathogens at the titres outlined as unsatisfactory in guidelines from the Health Protection Agency (now Public Health England) (2009). Thus, the NGS methods adopted were not sensitive enough to use as a routine screening tool to ensure food safety. Similar conclusions have been drawn from other studies endeavouring to use NGS as a tool to screen other food matrices for pathogens. For example, Leonard *et al.* (2015) found that without pre-enrichment *Escherichia coli* spiked into spinach was only

detectable at levels of 10^7 CFU/sample. However, de Boer *et al.* (2015), report a more sensitive LoD (2 x 10^2 spores ml^{-1}) for *Bacillus* spp. in canned soup utilising 454-pyrosequencing, and Frey *et al.* (2014) report a LoD of ≈1 x $10^{2.5}$ PFU ml^{-1} for dengue virus in blood using 454-pyrosequencing. The relative insensitivity of the approaches tested explains the lack of capability to detect low level norovirus and mengo virus without enrichment.

Culture-based enrichment methods may improve the sensitivity of NGS approaches for bacterial pathogens and viruses (Matias *et al.* 2010; Leonard *et al.* 2015; Rosimin *et al.* 2016), but also lead to a significant change in the microbiome composition and will affect the LoD of non-bacterial targets (Jarvis *et al.* 2015; Hyeon *et al.* 2017). Enrichment therefore negates the benefits of using a non-targeted approach, including the potential for source tracking or hygiene monitoring through the comparison of the microbiomes associated with samples and food production processes (Newton *et al.* 2013; Bartsch *et al.* 2018).

ScriptSeq sequencing yielded a more sensitive LoD than NEBNext, possibly because although both methods use ribosomal depletion to enrich the samples, the ScriptSeq method removes chloroplast RNA, in addition to mitochondrial and cytoplasmic rRNA, and thus problems caused by contamination with host RNA are smaller. One of the most notable findings was that the ScriptSeq kit delivered greater sensitivity than 16S rRNA gene amplicon sequencing, which due to its semi-targeted nature is widely regarded as the most sensitive technique. The results presented show that, as previously described in the literature (Hanshew *et al.* 2013; Beckers *et al.* 2016), the universality of the 16S rRNA gene primers leads to cross-amplification of the 16S rRNA gene from the host chloroplasts, due to its similarity in sequence to the bacterial 16S rRNA gene.

There were large differences in the LoD observed between bioinformatics methods for assignment of *Salmonella*. This is largely due to the incorrect assignment of non-*Salmonella* bacteria and chloroplast RNA to conserved regions of the *Salmonella* genome, notably 16S and 23S rRNA genes, seen for mapping to the reference using BWA, assignment using Kraken with the full database for metatranscriptomics analysis, and both QIIME and QIIME2 for 16S rRNA gene analysis. This leads to a lack of correlation between the input concentration of *Salmonella* and the number of assigned reads and therefore these bioinformatic methods should be avoided in this context due to the potential for false positives.

A more sensitive LoD may have been achieved had there been greater sequencing depth. Within replicates of the same titre of spiked MS2 or *Salmonella*, much of the variation in reads of MS2 or *Salmonella* correlated with the total number of reads obtained for that sample, showing the effects of sequencing depth on sensitivity. The sequencing depths chosen for these experiments reflected realistic uses of the techniques, with amplicon sequencing being used at considerably lower depths than metatranscriptomics and was highlighted in this study through the use of read numbers not normalised reads or percentages. The sequencing depth could be increased through the running of fewer samples on a single MiSeq run, although this would make it prohibitively expensive for routine use, or through the use of alternate platforms such as the HiSeq or NovaSeq, although little has been published in the literature on the effect of the shorter read lengths of these platforms on the LoD.

To examine this effect a comparison was run on a simulated dataset with increasing levels of reads associated with *Salmonella* to mimic the LoD run. The samples were then analysed using the most sensitive technique identified in the LoD run – Kraken with the mini database – to allow for the direct comparison of outputs from these three platforms. The finding that all these platforms give no significant difference in the detection of *Salmonella* at equal read depth is a key finding and allows for the use of these platforms, which are cheaper per sample than the MiSeq, without loss of sensitivity at the same depth of sequencing. This makes it financially achievable to sequence at a greater depth. The more sensitive LoD established for this method when compared to the LoD found in the lab may be due to the decreased levels of "host contamination" specified for the mock community, thereby increasing the percentage of reads associated with the microbiome. This again shows that removal of the host RNA is an important step in the sequencing of the microbiome and affects the LoD.

2.5 Conclusions

The identified LoD on the MiSeq is not sensitive enough to detect low level pathogen contamination in a biologically complex fresh produce sample without incorporating an enrichment step in the sample preparation protocol. Therefore, current NGS technologies are not suitable for non-targeted routine screening of fresh produce for human pathogens. HiSeq and NovaSeq simulated reads show that the LoD of these short read technologies do not affect the LoD identified, and therefore the increased depth per sample cost of these

platforms may deliver an increase in sensitivity. This is the first study reporting the LoD of current sequencing and bioinformatics approaches as applicable to the analysis of microbial contamination of fresh produce. This study should be treated as the first step in a validation procedure looking at the use of NGS within the fresh produce microbiome and informs on the key drawbacks of the currently available methods. Future studies should focus on increasing the sensitivity of NGS, for example through increased depth of sequencing, to obtain a more representative measure of the microbiome.

Chapter 3. Metatranscriptomics of the fresh produce microbiome

3.1 Introduction

Current pathogen screening methodologies are labour-intensive, overly-prescriptive, slow to deliver results and expensive when conducted in bulk, as they are inevitably performed to service due diligence requirements applicable to determining the safety of fresh produce for human consumption (Aw *et al.* 2016). To ensure due diligence, current testing methodologies often also test for indicator organisms. Indicator organisms are present at higher concentrations than those of human pathogens and are used to reveal problems within the production chain (Aguado *et al.* 2004; Health Protection Agency (now Public Health England) 2009). Their presence may indicate faecal contamination events, poor process control, and/or the potential for contamination with human pathogens. The Health Protection Agency (now Public Health England) (2009) Guidelines for Assessing the Microbiological Safety of Ready-to-Eat (RTE) Foods list Enterobacteriaceae, *Escherichia coli* and *Listeria* species as hygiene indicator organisms for RTE – a food category that includes fresh produce and leafy greens. Both Enterobacteriaceae and *E. coli* often originate in the animal and human intestinal tract but are also found in plant and soil microbiomes. *Listeria* spp. originate in the environment, survive on processing equipment and show a capability to multiply and grow at 4^{o}C. A significant issue with the application of these organisms is that many studies reveal no correlation between currently used indicator organisms and the presence of human pathogen contamination of food or outbreaks of food poisoning (Wells and Butterfield 1997; Harwood *et al.* 2014; Orlofsky *et al.* 2016).

Although the limit of detection of Next Generation Sequencing (NGS) technologies is insufficient to detect trace level contamination of fresh produce, as shown in Chapter 2, there is a lot of other information we can gain from NGS techniques. For example, the presence of antimicrobial resistance (AMR) associated genes, the composition of the microbiome, and, in the case of metatranscriptomics, the actively transcribing genes within, and thereby function of, the microbiome. If NGS microbiome data is combined with information from traditional methodologies on the presence of human pathogens, the resulting data may be used to identify novel indicator organisms, or profiles of multiple organisms, that better correlate with contamination of food with human pathogens, or potential members of the microbiome that encourage the survival of human pathogens in

the microbiome. Previous studies have shown that the presence of certain species within the microbiome can correlate with the survival of human pathogens. It has been shown that presence of soft rot pathogens, such as *Pectobacterium,* correlates with enhanced *Salmonella* survival on fresh produce (Wells and Butterfield 1997). This may be due to the nutrient-rich environment associated with the damage caused by soft rot which encourages *Salmonella* to flourish and outcompete other elements of the microbiome (Fatica and Schneider 2011; Deering *et al.* 2012). Utilising NGS may also help identify elements of the microbiome that negatively correlate with human pathogens. This may facilitate the discovery of novel biocontrol agents. Heaton and Jones (2008b), for example, found that growth of *E. coli* may be suppressed by the presence of certain epiphytic bacteria, for example some *Enterobacter* and *Pseudomonas* species.

The microbiome also affects the survival of human pathogens through the production of antimicrobial compounds (Mendes *et al.* 2013). Many members of the fresh produce microbiome produce antimicrobial compounds, for example *Leuconostoc* spp. (Trias *et al.* 2008a). The antimicrobials produced by these species have been proven to influence the survival of human pathogens within the fresh produce microbiome (Trias *et al.* 2008b; Olaimat and Holley 2012). In addition, the presence of antimicrobials produced by these organisms may also create a selection pressure leading to the increased acquisition of AMR associated genes (Martínez *et al.* 2007), thereby increasing incidence of AMR bacteria in the fresh produce microbiome. Other factors that may influence the abundance of AMR bacteria in the fresh produce microbiome include the supplementation of the soil with manure from animals treated with antibiotics (Marti *et al.* 2013), or the use of non-potable water for irrigation (Cerqueira *et al.* 2019). Current research on the prevalence of AMR genes within the fresh produce microbiome is limited and therefore this is a key area of study to assess the impact of AMR within the fresh produce supply chain.

The aims of this study were to:

(i) Characterise the actively transcribed microbiome associated with commercial samples of fresh produce.
(ii) Identify active members of the microbiome that correlate with human pathogen contamination and thus afford potential as indicators of contamination events or pathogen suppression.

(iii) Determine the prevalence of AMR-associated genes in the microbiome associated with fresh produce.

3.2 Materials and Methods

3.2.1 Sample handling and receipt

Samples of fresh produce (full details in Table 10) were collected during routine production at a major sandwich production facility.

The first step on the production line was the washing of produce. Washing was undertaken by one of two methods dependent on produce type (as outlined in Table 9). Iceberg Lettuce, Baby Spinach, Cos Lettuce, Apollo Lettuce and Rocket were washed using the Kronen Washer, a two-stage continuous salad washer. An initial washer stage was undertaken using untreated water wash (mains water) and followed by a second wash stage within a chemical dosed washer (chilled water) which was monitored by a Prominent System (an automated chemical dosing system) to ensure a concentration of free chlorine was maintained between >40ppm-<80ppm, and citric acid used to achieve a pH between 6.0-8.0 to allow for optimum presence of free chlorine. The flow and submersion of ingredients through both washers was controlled by a series of water jets. All other produce was washed using the Atir washer, a single stage batch washer. Produce were washed using a single chemical wash, using a chilled water supply. The chemical wash was monitored by a Prominent System as described in the chemical wash stage of the Kronen Washer protocol. The length of wash was product dependent (Table 9) and timed with a digital clock / stopwatch to ensure accuracy.

Table 9. Chemical wash details for produce included in this chapter, A) shows products washed on the Kronen washer, B) shows products washed using the Atir washer.

A

Product	Max. loading and wash rate (kg per minute)	Target pH Level	Target Free Chlorine Level
Iceberg Lettuce	6	6.0 – 7.5	>40ppm-<80ppm
Baby Spinach	4	6.0 – 7.5	>40ppm-<80ppm
Cos Lettuce	4	6.0 – 7.5	>40ppm-<80ppm
Apollo Lettuce	4	6.0 – 7.5	>40ppm-<80ppm
Rocket	4	6.0 – 7.5	>40ppm-<80ppm

B

Product	Max. single load (kg)	Wash time (mins)	Target pH Level	Target Free Chlorine Level
Onions Diced	10	10	6.0 – 8.0	>40ppm-<80ppm
Red Onion Sliced	10	10	6.0 – 8.0	>40ppm-<80ppm
Red Onion Diced	10	10	6.0 – 8.0	>40ppm-<80ppm
Cress	10	7	6.0 – 8.0	>40ppm-<80ppm
Coriander	4	7	6.0 – 8.0	>40ppm-<80ppm
Baby Watercress	5	10	6.0 – 8.0	>40ppm-<80ppm
Spring Onions Sliced	15	9	6.0 – 8.0	>40ppm-<80ppm

Samples were collected post wash, split in to two, and one half sent *via* refrigerated same-day transport to a contract lab to undergo routine microbiological testing, the other half was sent to Fera *via* next day delivery in a refrigerated vehicle. Upon receipt at Fera samples were immediately placed in -80°C storage until analysis.

3.2.2 Sample extraction

Samples were taken from -80°C storage and powdered while still frozen to allow the subsampling of a mixture of parts of each sample. A 50 mg sample of frozen tissue was placed in a pestle and mortar and ground to a homogenous paste. To each sample, 180 µl of buffer RLT + β- Mercaptoethanol was added and then pipetted into a QIAshredder spin column (Qiagen, Hilden, Germany) and centrifuged at 8000 *g for 1 minute to homogenise. The supernatant was then combined with 0.5 volume of ethanol and mixed, before extracting RNA using the RNeasy plant mini kit (Qiagen, Hilden, Germany) following the manufacturer's protocol (Qiagen 2012b) "Purification of total RNA from plant cells and tissues" with inclusion of on-column DNase digestion, eluting in 2 x 50 µl molecular biological grade water (MBGW). The eluate was transferred to a new 1.5 ml tube and stored at -80 °C.

3.2.3 Sequencing

Samples of RNA extract were quantified using the Nanodrop 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, United States), then diluted to achieve a concentration of 1 µg RNA in 10 µl MBGW. Preparation for sequencing was undertaken using the TruSeq Stranded Total RNA with Ribo-Zero Plant kit (Illumina, San Diego, United States) following the manufacturer's instructions (Illumina 2018d). Samples were quantified using the Nanodrop 2000 and diluted to less than 1 µg total RNA in 10 µl. To each sample, 5 µl of rRNA binding buffer and 5 µl rRNA removal mix (plant) were added and the mix incubated at 68 °C for 5 mins, held at 4 °C, then incubated at RT for 1 min. To each sample, 35 µl of rRNA removal beads were added, incubated at RT for 1 min and placed on a magnetic stand for 1 min, and the supernatant transferred to a new plate. Samples were cleaned using Ampure XP beads (Appendix A), eluting in 8.5 µl elution buffer. To the supernatant, 8.5 µl elution, bind, fragment high mix was added and incubated at 94 °C for 8 mins and hold at 4 °C. A master mix containing 1:9 ratios of Superscript II:first strand synthesis mix was made and 8 µl added to each sample, then the mix incubated at 25 °C for 10 mins, 42 °C for 15 mins, 70 °C for 15 mins, then the temperature decreased to 4 °C. After PCR, 10 µl resuspension buffer

and 20 µl 2nd strand marking master mix were added to each sample and incubated at 16 °C for 1 hour before leaving at RT to warm. Samples were cleaned using Ampure XP beads (Appendix A), eluting in 15 µl resuspension buffer. An additional 2.5 µl resuspension buffer was added followed by 12.5 µl A-tailing mix and incubated at 37 °C for 30 mins, 70 °C for 5 mins, then the temperature decreased to 4 °C. To each sample, 2.5 µl resuspension buffer, 2.5 µl ligation mix, 2.5 µl RNA adapters were added and incubated at 30 °C for 10 mins, then the temperature decreased to 4 °C. To each sample, stop ligation buffer was added to stop the activity of the ligase. Samples were then cleaned twice using Ampure XP beads (Appendix A), eluting in 20 µl resuspension buffer. To each sample, 5 µl PCR primer cocktail and 25 µl PCR master mix were added and thermal cycled for 98 °C for 30 s, followed by 15 cycles of 98 °C for 10 s, 60 °C for 30 s and 72 °C for 30 s, before a final extension of 72 °C for 300 s. Samples were finally cleaned using Ampure XP beads, eluting in 30 µl resuspension buffer

The concentration of sequencing ready DNA was quantified using the Qubit® DNA HS Assay Kit (Life Technologies 2015a) and pooled at equimolar concentrations. Free adapters were removed from the pooled samples to prevent index hopping using Illumina Free Adapter Blocking Reagent (Illumina, San Diego, United States), and the processed pool containing sequencing ready DNA was quantified using the Qubit® DNA HS Assay Kit (Life Technologies 2015a) and quality checked using the Agilent 2200 TapeStation system with High Sensitivity D1000 reagents (Agilent Technologies 2015). The sample peak was between 200 bp and 1000 bp, with no small fragments below 200 bp, thus meeting the required criteria for HiSeq analysis. The pool was frozen at -80 °C and transferred to Leeds Institute of Molecular Medicine NGS facility (http://dna.leeds.ac.uk/genomics/hiseq.php) and run on a single lane of the HiSeq 3000 creating 2x150 bp reads.

3.2.4 Quality Control

Extraction blanks were undertaken as part of the RNA extraction. Process blanks, MBGW put through the same processing as samples, and indexing blanks, MBGW added at indexing PCR stage, were undertaken for each sequencing method. All were examined using the tapestation and Qubit for quality purposes. Blanks were run on the sequencer as a separately indexed sample. All samples, including blanks, were examined for read number and those with low quality reads, or less than 500 reads were filtered out of the analysis. Any blanks remaining were run through the bioinformatic analysis separately to samples and

the top taxa compared manually to those in experimental samples to rule out cross contamination.

Due to being run on the HiSeq at Leeds Institute of Molecular Medicine NGS facility, the final pool was prepared to their specification, and run metrics, including the use of the PhiX internal standard, mapped to the PhiX genome as part of the standard Illumina workflow run quality, in addition to metrics on cluster density and read numbers was assessed as part of their service.

3.2.5 Bioinformatics

Reads from the HiSeq were trimmed using Sickle v1.33 (Joshi and Fass 2011) to remove sequence of quality less than Q20 (1 in 100 probability of incorrect base call) and lengths less than 100 base pairs. Data were then subject to two different paths of analysis (Figure 8). Initial analysis was undertaken to identify AMR associated genes, comprising of assembly of the trimmed reads using Trinity v2.8.4 (Haas *et al.* 2013) followed by screening of AMR genes using ABRicate v0.8 (Seemann 2018). All genes with more than 80% identity or 80% coverage were recorded. From the trimmed reads, the plant matrix genome was removed by mapping the reads to the matrix genome using BWA-MEM 0.7.17 (Li and Durbin 2009). Mapping was performed against the full genome or, when unavailable, the matrix mitochondrial genome, or else where no genetic information was available, against the RefSeq plant chloroplast database (see Table 10). The data were then filtered to remove data associated with the matrix genome using Samtools v1.9 (Li *et al.* 2009) and a fastq file created for each sample using Bedtools v2.280 (Quinlan and Hall 2010) to obtain files containing the unmapped reads that were able to undergo taxonomic assignment. The taxonomy was assigned to unmapped reads using Kraken v1.1 using the MiniKraken database (Wood and Salzberg 2014) and a table of taxonomy *versus* read numbers was made and filtered to remove taxa only present in a single sample. This was then used in conjunction with metadata as an input in to Lefse (Segata *et al.* 2011; Huttenhower 2019) to examine differential taxa abundance based on the metadata associated with the sample (Table 10). Briefly, Lefse performs a Kruskall-Wallis test to analyse all taxa, testing whether the abundances in different metadata classes are differentially distributed. If any taxa violate the null hypothesis, Lefse then runs a pairwise Wilcoxon test to check whether all pairwise comparisons between samples within different metadata classes significantly agree with the class level trend. The resulting values are used to build a Linear Discriminant Analysis model

which ranks the taxa association with each metadata class. This is then used to plot significant ($p \leq 0.05$) results. The metadata used as input for Lefse were (as outlined in Table 10 and Table 11): Enterobacteriaceae positive or negative microbiology, *Listeria* spp. positive or negative microbiology, produce type, and the presence and absence of AMR associated genes based on outputs from ABRicate. The outputs were manually interrogated for biological and statistical relevance.

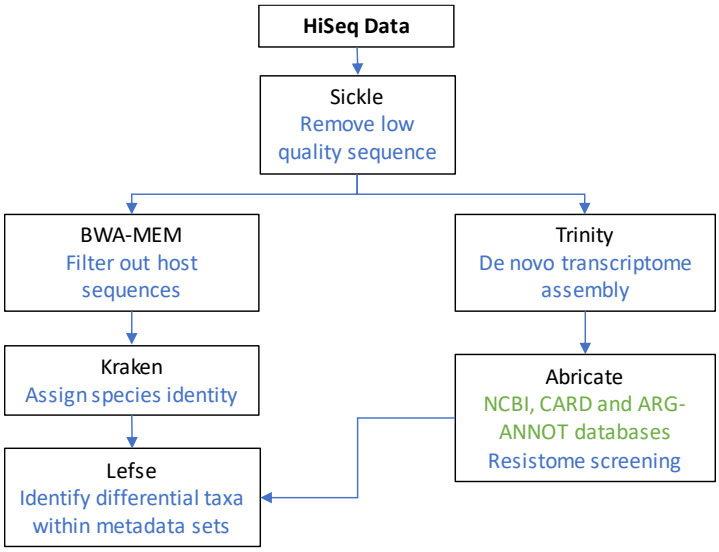


Figure 8. Bioinformatics methodologies employed to examine AMR associated genes and identify differential taxa within metadata sets.

3.3 Results

3.3.1 Quality Assessment

Negative controls were examined, and those with read numbers of less than 500, were filtered out at QC stage and no subsequent analysis was performed on them. Negative controls which did not fall below this threshold (likely in part due to the higher depth of sequencing) were run through same analysis pipelines as experimental samples and examined to check for evidence of contamination. The top taxa did not overlap between samples and controls, suggesting no contamination, but no firm thresholds exist as this is a developing field.

3.3.2 Microbiological Testing Results

Microbiological testing results provided by Westward Laboratories are summarised in Table 10. Enterobacteriaceae results below the laboratory's standard dilution-series detection

limit (<100 colonies) were classified as Not Detected. Those above the detection limit were classified as Detected.

Table 10. Fresh produce sample details; type, subcategory, presence/absence data for Enterobacteriaceae and *Listeria* spp. and genome employed for subtraction in NGS analysis.

Item Description	Category	Enterobact-eriaceae	*Listeria* spp.	Genome Subtracted
Baby Watercress	Leafy Green	Not detected	Not detected	*Nasturtium officinale* chloroplast
Apollo	Leafy Green	Not detected	Not detected	*Lactuca sativa* genome
Baby Spinach	Leafy Green	Not detected	Not detected	*Spinacia oleracea* full genome
Red Onion Diced	Onion	Detected	Not detected	*Allium cepa* chloroplast
Cress	Leafy Green	Detected	Not detected	Refseq plant chloroplast database
Iceberg Lettuce	Leafy Green	Not detected	Not detected	*Lactuca sativa* genome
Cos Lettuce	Leafy Green	Not detected	Not detected	*Lactuca sativa* genome
Spring Onions Sliced	Spring onion	Detected	Not detected	*Allium cepa* chloroplast
Corriander	Leafy Green	Not detected	Not detected	*Coriandrum sativum* chloroplast
Baby Spinach	Leafy Green	Detected	Not detected	*Spinacia oleracea* full genome
Red Onion Sliced	Onion	Detected	Detected	*Allium cepa* chloroplast
Red Onion Diced	Onion	Detected	Not detected	*Allium cepa* chloroplast
Spring Onions Sliced	Spring onion	Detected	Not detected	*Allium cepa* chloroplast
Coriander	Leafy Green	Detected	Not detected	*Coriandrum sativum* chloroplast
Red Onion Diced	Onion	Detected	Not detected	*Allium cepa* chloroplast
Coriander	Leafy Green	Not detected	Not detected	*Coriandrum sativum* chloroplast
Cos lettuce	Leafy Green	Not detected	Not Detected	*Lactuca sativa* genome
Baby Watercress	Leafy Green	Detected	Not Detected	*Nasturtium officinale* chloroplast
Iceberg	Leafy Green	Not detected	Not Detected	*Lactuca sativa* genome
Apollo	Leafy Green	Not detected	Not Detected	*Lactuca sativa* genome
Baby Spinach	Leafy Green	Detected	Not Detected	*Spinacia oleracea* full genome
Red Onion Diced	Onion	Detected	Not Detected	*Allium cepa* chloroplast
Coriander	Leafy Green	Detected	Not Detected	*Coriandrum sativum* chloroplast
Cress	Leafy Green	Detected	Not Detected	Refseq plant chloroplast database
Red Onion Diced	Onion	Detected	Detected	*Allium cepa* chloroplast
Baby Spinach	Leafy Green	Not detected	Not Detected	*Spinacia oleracea* full genome
Red Onion Sliced	Onion	Detected	Detected	*Allium cepa* chloroplast
Spring Onions Sliced	Spring onion	Detected	Not Detected	*Allium cepa* chloroplast
Baby Watercress	Leafy Green	Not detected	Not Detected	*Nasturtium officinale* chloroplast
Coriander	Leafy Green	Detected	Not Detected	*Coriandrum sativum* chloroplast
Cress	Leafy Green	Detected	Not Detected	Refseq plant chloroplast database
Onions Diced	Onion	Detected	Not Detected	*Allium cepa* chloroplast
Coriander	Leafy Green	Detected	Not Detected	*Coriandrum sativum* chloroplast
Baby Watercress	Leafy Green	Detected	Detected	*Nasturtium officinale* chloroplast
Spring Onions Sliced	Spring onion	Detected	Not Detected	*Allium cepa* chloroplast
Red Onion Sliced	Onion	Detected	Not Detected	*Allium cepa* chloroplast
Red Onion Diced	Onion	Detected	Not Detected	*Allium cepa* chloroplast
Baby Spinach	Leafy Green	Detected	Not Detected	*Spinacia oleracea* full genome
Cress	Leafy Green	Detected	Not Detected	Refseq plant chloroplast database
Apollo	Leafy Green	Not detected	Not Detected	*Lactuca sativa* genome
Baby Watercress	Leafy Green	Detected	Not Detected	*Nasturtium officinale* chloroplast
Onions Diced	Onion	Detected	Not Detected	*Allium cepa* chloroplast
Red Onion Sliced	Onion	Detected	Detected	*Allium cepa* chloroplast
Coriander	Leafy Green	Detected	Not Detected	*Coriandrum sativum* chloroplast

3.3.2 Sequencing Results

Between 4,387,110 and 55,790,718 reads were produced for samples that passed HiSeq QC, with an average of 15,709,445 reads (Table 11). Quality was checked using Sickle, taking the average read number down to 15,268,388 reads. The number of reads assigned to a taxa in the Kraken database was on average 196,276, although this varied widely between samples with the lowest being 678 reads, and the highest 1,281,892 reads (Table 11). This equates to the percentage of reads assigned to taxa of between 0.02 - 5.3% of reads per sample, with the average being 1.3% of reads assigned to taxa.

A total of 56 unique AMR associated genes were identified across 23 of 53 samples. Most were of low identity and coverage. When the data were filtered to include only genes with greater than 80% identity and 80% coverage, four genes (*CRP, H-NS, MexF, MexB*) were identified across seven samples. Of the seven samples yielding over 80% identity and coverage, there were three samples which contained multiple AMR associated genes, two containing both *CRP* and *H-NS*, and one containing both *MexF* and *MexB*. Of these seven samples, five were from onion (out of 12 samples) and two were from leafy greens (out of 34 samples).

The top genera assigned across all samples were *Alteromonas*, *Pseudomonas*, *Leuconostoc* and *Rahnella* (average and total read numbers summarised in Table 12). All other genera identified had less than an average of 1000 reads per sample. The microbiome varied by produce sub-category (Figure 9). When samples were split by category of produce (as shown in Table 10), all produce types returned *Alteromonas* as the genus with the highest number of reads, but the other top members of the microbiome varied (Table 10). When Lefse was used to identify differences in the fresh produce microbiome associated with category of produce, the greatest number of genera correlated with onions (11), three genera correlated with spring onions and one genus correlated with leafy greens (Figure 10 A).

Table 11. Read numbers per sample; from the HiSeq, after QC using trinity, post host subtraction, and reads assigned by Kraken. Presence of AMR genes and whether they had >80% identity and coverage, Enterobacteriaceae and *Listeria* read counts per sample, colour coded by consistency with microbiological results (read no. greater than 50) - blue false positives, green correct positives, orange false negatives, white true negatives.

Sample name	Read no. from HiSeq	Read no. post QC	Read no. post host subtraction	Reads assigned by Kraken	AMR Gene Found	AMR genes over 80% coverage / identity	Enterobac-teriaceae read no.	*Listeria* spp. read no.
302850_4_5	26089266	25431276	14173687	189015	N	N	74	0
302907_4_5	11892030	11576818	128060	16706	N	N	17	0
302910_4_5	10232472	9908688	44518	2749	N	N	13	0
303082_4_5	5099030	4966640	3866574	129535	Y	Y	20698	8
303124_4_5	34277818	33458770	19362069	570698	Y	Y	65871	2
303187_4_5	7118490	6910946	115510	7490	N	N	25	0
303221_4_5	7342508	7122400	77186	15925	N	N	11	0
303239_4_5	50973622	49669690	40787193	326709	Y	N	510	456
303245_4_5	7945194	7756306	4181242	59475	N	N	7	0
308845_14_6	4387110	4252440	13706	678	N	N	2	0
309228_14_6	5443568	5305152	4359199	266347	Y	Y	169160	5
309229_14_6	12935072	12682002	8359466	162898	Y	Y	26091	2
309230_14_6	16664670	16240822	9556147	124083	N	N	645	58
309238_14_6	11968696	11594144	5184573	65079	N	N	14	0
313154_12_7	12904162	12395304	11571374	241124	Y	N	6038	64
313166_12_7	17365440	16764826	7447240	128613	N	N	17	0
318663_23_8	8073844	7889002	133120	8597	N	N	135	0
318982_23_8	6412722	6232270	3995802	204533	Y	N	2501	2
319002_23_8	11468056	11099836	213057	13502	N	N	171	6
319003_23_9	11095966	10820782	137393	12257	N	N	5	0
319004_23_8	6163996	5986374	72495	6611	N	N	296	0
319093_23_8	12709418	12371028	10438651	277037	Y	N	25004	21
319100_23_8	17880148	16786130	11789545	390401	Y	N	1492	0

(Table 10 continued on next page)

Table 11. Read numbers per sample; from the HiSeq, after QC using trinity, post host subtraction, and reads assigned by Kraken. Presence of AMR genes and whether they had >80% identity and coverage, Enterobacteriaceae and *Listeria* read counts per sample, colour coded by consistency with microbiological results (read no. greater than 50) - blue false positives, green correct positives, orange false negatives, white true negatives.(continued)

Sample name	Read no. from HiSeq	Read no. post QC	Read no. post host subtraction	Reads assigned by Kraken	AMR Gene Found	AMR genes over 80% coverage / identity	Enterobac-teriaceae read no.	*Listeria* spp. read no.
320046_13_9	20884458	20286496	11202628	189452	Y	N	2458	0
321909_13_9	7858428	7622666	6568444	131712	Y	N	19777	35
321994_13_9	26889614	26193086	132273	7182	N	N	398	0
322066_13_9	9523120	9348218	6078696	302483	Y	Y	156850	1
322068_13_9	9473536	9280728	6631304	81237	Y	N	919	65
322162_13_9	5602436	5490016	3251805	60474	N	N	152	0
322236_13_9	20665710	20205154	12035278	186406	N	N	34	0
326777_18_10	40637404	39666532	22755815	521226	Y	N	3545	0
326945_18_10	10678624	10409018	8467217	399557	Y	N	222	2
326953_18_10	14791476	14419110	7088068	120467	Y	N	336	0
326996_18_10	18640936	18102172	10418615	679687	Y	Y	11036	2
327019_18_10	55790718	54259102	7249593	199229	N	N	6320	4
327089_18_10A	24204234	23256192	33775122	1281892	Y	N	356	26
327089_18_10B	9839550	9609030	14580038	341004	Y	N	47862	6
331896_21_11	8824026	8568764	36115	9475	N	N	25	0
331962_23_11	8331798	8097694	4475831	91732	Y	N	2847	0
332045_23_11	8359864	8144074	101945	15569	N	N	9	0
332105_23_11	23850566	23114328	11816301	284279	N	N	114	0
332109_23_11	19090362	18687200	15701686	292078	Y	Y	71375	16
332110_23_11	14336984	13836284	11663132	203459	Y	N	13027	11
332117_23_11	16498456	15991586	6400778	138483	Y	N	55	0

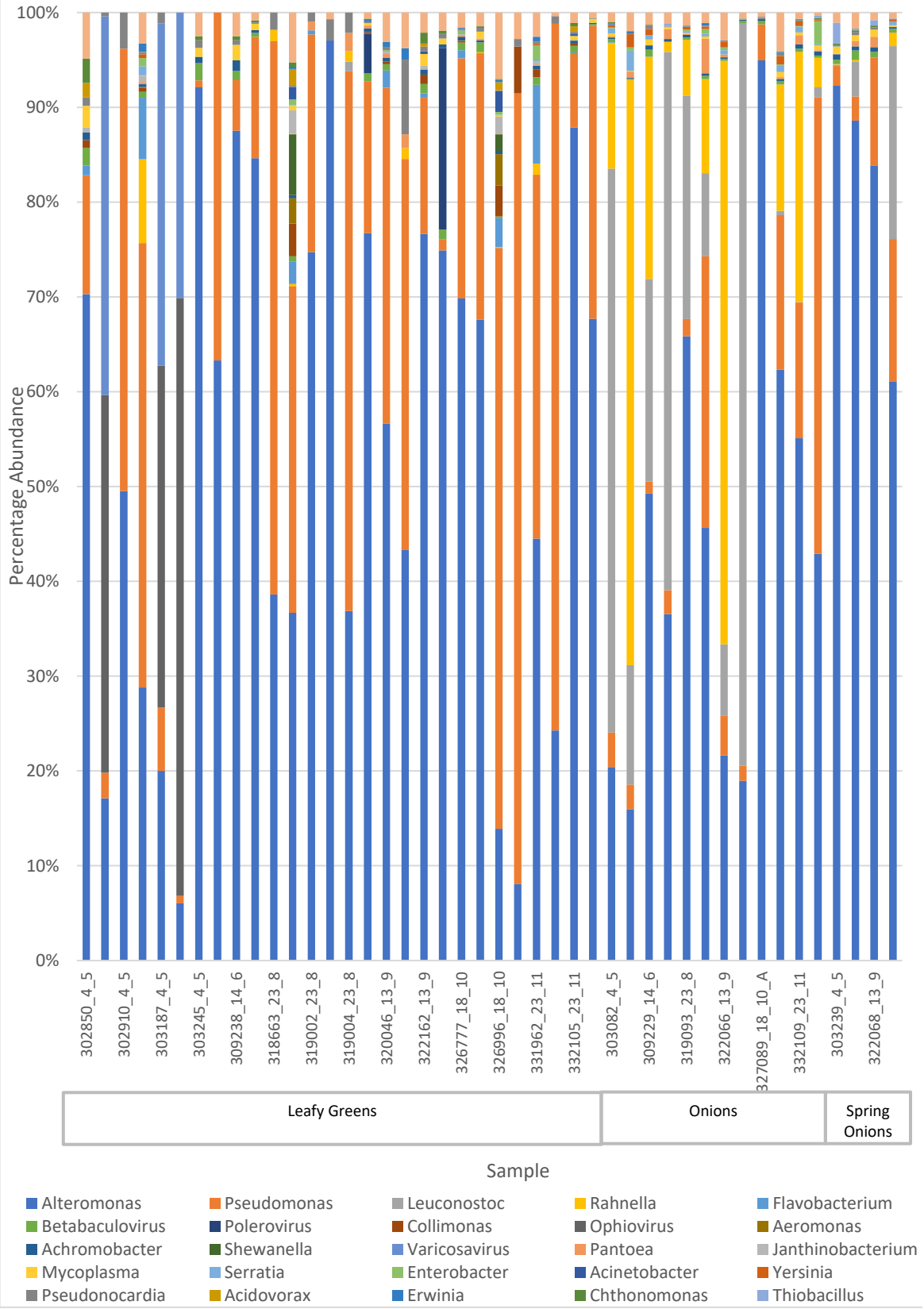


Figure 9. Graph showing percentage abundance for top 25 genera detected using Kraken in samples analysed by HiSeq metatranscriptomics.

Table 12. Table showing the top genera identified for all samples, fresh produce, onions and spring onions, and the total and average number of reads assigned to each genus.

Samples	Genus	Total Read Number	Average Read Number
All Samples	*Alteromonas*	3031644	57201
	Pseudomonas	928560	17520
	Leuconostoc	574317	10836
	Rahnella	370050	6982
Fresh Produce	*Alteromonas*	1092971	32146
	Pseudomonas	686039	20178
Onion	*Alteromonas*	1573897	104926
	Leuconostoc	543673	36245
	Rahnella	341637	22776
	Pseudomonas	211132	14075
Spring Onion	*Alteromonas*	364776	91194
	Pseudomonas	31389	7847
	Leuconostoc	30520	7630

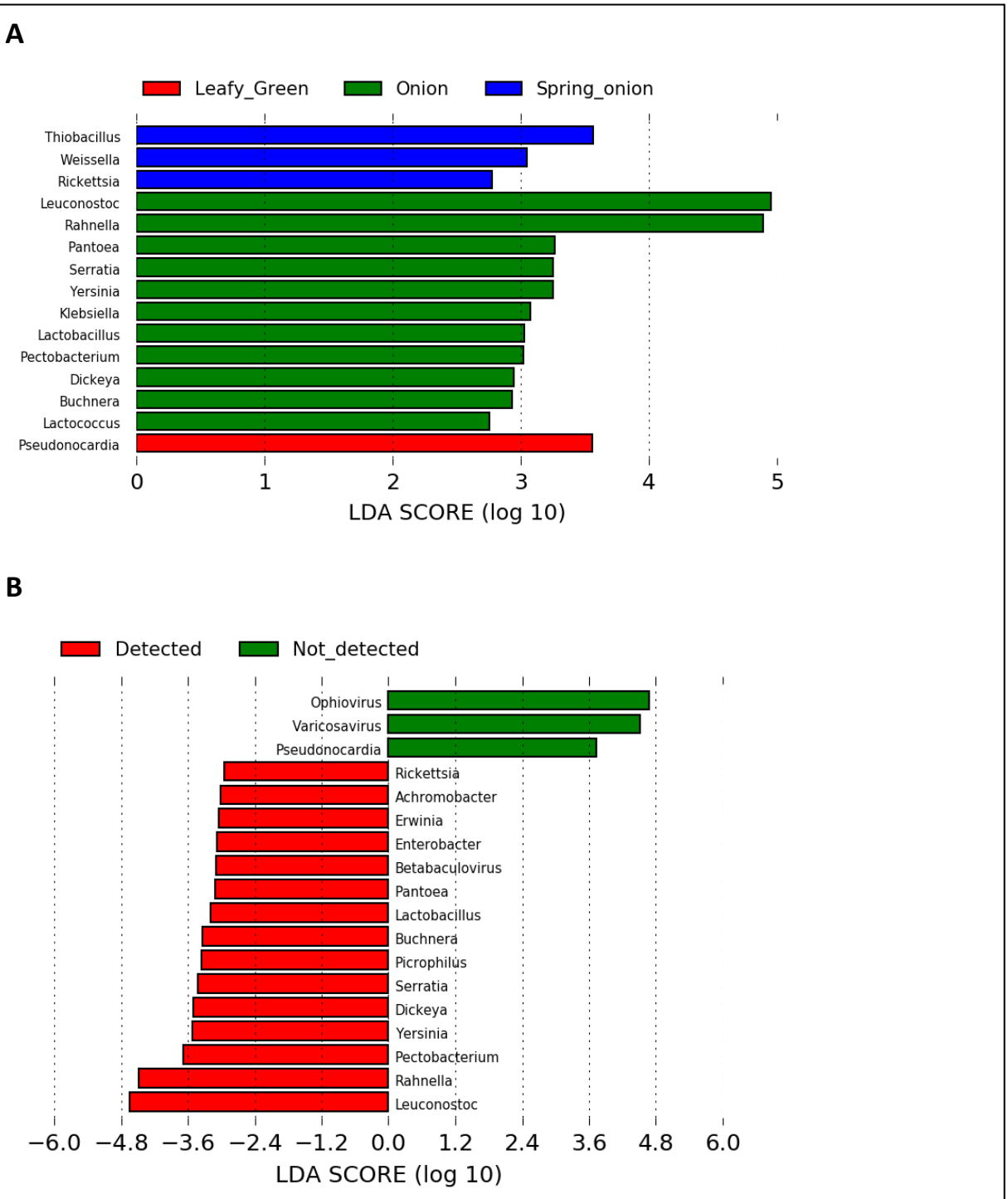
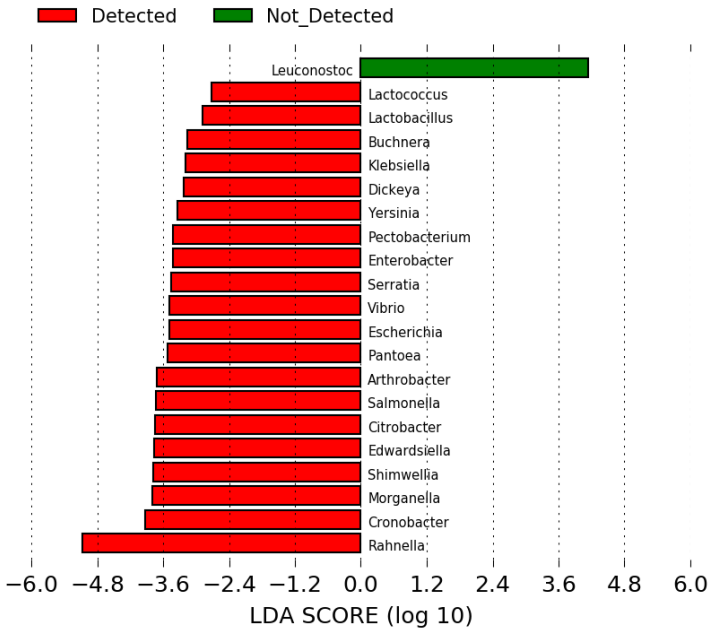


Figure 10 (continues on next page)

C



D

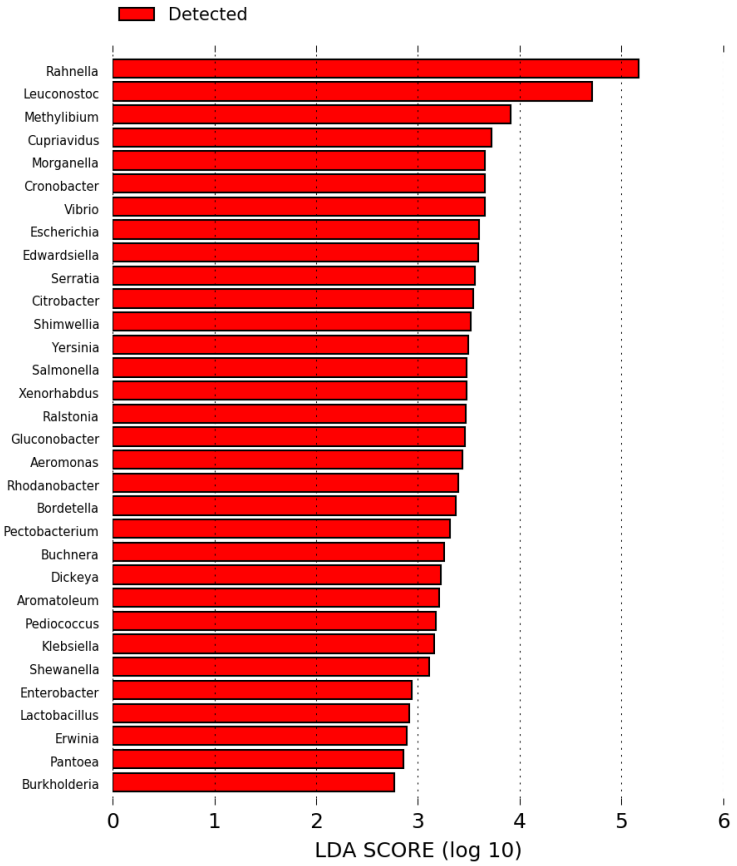


Figure 10. Graphs showing differential bacterial abundances at genus level for metadata associated with fresh produce samples calculated using Lefse. A: differential abundance for samples split by produce types: leafy greens, onions, spring onions; B: differential abundance for samples split by positive or negative for Enterobacteriacea; C: differential abundance for samples split by positive or negative for *Listeria* spp.; D: differential abundance for samples split by positive or negative for presence of antimicrobial resistance genes.

From Kraken taxonomy assignments, all samples that passed HiSeq QC (44) had reads assigned to Enterobacteriaceae (Table 11). Of these, 32 samples exhibited more than 50 reads; 27 of which equated to positive Enterobacteriaceae in microbiological testing, and 5 of which equated to samples where Enterobacteriaceae had remained undetected in routine microbiological testing. Four samples positive by microbiological testing for Enterobacteriaceae exhibited read numbers of less than 50. For *Listeria* spp., all five samples that tested positive in routine microbiological testing and passed HiSeq QC had reads assigned to *Listeria*, although none of these had greater than 50 reads. Interestingly, a further 15 samples also had reads assigned to *Listeria*, four of which had over 50 reads, despite not being found to test positive for *Listeria* spp. using conventional microbiology (Table 11).

Lefse examination of the microbiome in combination with microbiological data on presence or absence of Enterobacteriaceae (Table 10) identified 15 genera correlated with detectable levels of Enterobacteriaceae, and three genera correlated with no detectable Enterobacteriaceae (Figure 10 B). Of these, three genera were in fewer than 10 samples and 15 genera in 10 or greater (*Pseudonocardia*, *Achromobacter*, Betabaculovirus, *Pantoea*, *Picrophilus*, *Rahnella*, *Rickettsia*, *Erwinia*, *Leuconostoc*, *Serratia*, *Enterobacter*, *Yersinia*, *Lactobacillus*, *Dickeya* and *Pectobacterium*). *Pseudonocardia* was identified in the highest proportion of samples; detected in 41 samples.

For presence or absence of *Listeria* spp., 20 genera were correlated with detectable levels of *Listeria* spp., and one genus correlated with no detectable *Listeria* species (Figure 10 C). Of these, there were 12 genera in fewer than 10 samples and 9 genera in 10 or greater samples (*Pantoea*, *Rahnella*, *Dickeya*, *Leuconostoc*, *Serratia*, *Enterobacter*, *Yersinia*, *Lactobacillus* and *Pectobacterium*), with *Pantoea* found in the highest proportion of samples; detected in 25 samples.

There were 32 taxa positively correlated with samples containing AMR genes above 80% identity and coverage, and no taxa negatively correlated (Figure 10 D). Of these, there were 21 genera in fewer than 10 samples and 11 genera in 10 or greater samples (*Pantoea, Rahnella, Erwinia, Leuconostoc, Serratia, Enterobacter, Yersinia, Burkholderia, Lactobacillus, Dickeya* and *Pectobacterium*), with *Pantoea* again detected in the highest proportion of samples.

Nine taxa (*Rahnella, Pectobacterium, Yersinia, Dickeya, Serratia, Buchnera, Lactobacillus, Pantoea* and *Enterobacter*) were found which correlate with detection of Enterobacteriaceae*,* detection of *Listeria* spp. and the presence of AMR associated genes with greater than 80% identity. There were no taxa present negatively correlated with all metadata tested (Figure 10 B, C and D).

3.4 Discussion

The total burden of expressed AMR associated genes within the fresh produce microbiome was low, with only 15% (seven out of 44) of samples revealing AMR genes in this study. Few previous studies have looked at the incidence of AMR genes in the microbiome associated with leafy fresh produce, with these having given conflicting results depending on the methods utilised. Food associated bacteria screened for AMR phenotypes had a reported prevalence of 1% (Holzel *et al.* 2018). This is much lower than the rates identified by work in this thesis, and other studies utilising metagenomics. Metagenomics has reported higher levels of AMR-associated genes than both phenotypic screening and metatranscriptomics, with 20 classes of antibiotic resistance genes in the lettuce and radish microbiome (Fogler *et al.* 2019). The differences between the reported rates from the literature utilising metagenomics, and the work in this thesis utilising metatranscriptomics, may be due to the fact metagenomics will screen DNA and therefore all genetic information present within the microbiome, whereas metatranscriptomics will screen RNA and is therefore likely to only detect the actively transcribed genes within the microbiome, arguably the genes which are most important (Bashiardes *et al.* 2016).

In the current study, five of seven samples containing AMR associated genes originated from onions, with the other two being from leafy greens. Of the genes identified, *H-NS* was the most common (4 of 9 samples). This encodes a histone-like protein involved in global gene regulation including many membrane fusion protein and multidrug exporter genes. Homologues are found in several Gram-negative bacteria, notably *Escherichia coli*, *Klebsiella oxytoca* and several *Shigella* species (Nishino and Yamaguchi 2004; Jia *et al.* 2017b). The second most commonly detected gene was *CRP* (3 of 9 samples). This is another global regulator that affects the expression of the multidrug efflux pump MdtEF and has been characterised in the resistomes of numerous organisms, notably *Enterobacter* species, *E. coli*, *Salmonella enterica*, *Shigella* species and *Yersinia* species (Nishino *et al.* 2008; Jia *et al.* 2017a). The genes *mexF* and *mexB* were identified in a single sample; these both encode for

the inner membrane multidrug exporter of the efflux complex MexAB-OprM and are found on the same loci in *Pseudomonas aeruginosa* (Middlemiss and Poole 2004; Jia *et al.* 2017c, d). *Pseudomonas aeruginosa* was identified in the sample in which *mexF* and *mexB* were identified, suggesting this is the organism of origin of these genes. This is of notable importance as previous research has shown that *L. monocytogenes* may associate with *Pseudomonas* biofilms, encouraging survival and transfer of AMR associated genes to *L. monocytogenes,* thereby increasing the burden of AMR in human pathogenic organisms (Balcazar *et al.* 2015; Puga *et al.* 2018). Metatranscriptomic screens for AMR-associated genes are often unable to identify the organism of origin, and this study was unable to find the organism of origin for the other genes identified.

One of the limitations of the current study is that the analysis method employed to examine for AMR-associated genes only probed for genes believed to be directly associated with AMR, therefore may have missed AMR phenotypes caused by point mutations in antimicrobial targets. The identification of point mutations would be difficult to obtain from metagenomic data using currently available methodologies, due to the difficulty in determining which organism the genes are from and therefore the true sequence of the target gene. In addition, with low frequency genes it can be difficult to determine whether mutations are due to the presence of mutations within the organism or due to sequencing or analysis error. It is possible that additional information could be gained from this dataset by further examining the low identity or coverage matches in greater detail to identify functional mutations to antimicrobial targets or homologies to known AMR gene products. This approach might facilitate the identification of novel AMR associated genes (Adu-Oppong *et al.* 2016).

Testing with NGS found that 4/31 samples that were positive for Enterobacteriaceae using microbiology were negative using NGS, and none of the *Listeria* spp. positive samples (5) were detected using NGS. This observation maybe linked to the relative 'insensitivity' of NGS, with the LoD insufficient to detect low levels of contamination (see Chapter 2). To increase the sensitivity the depth of sequencing was increased, from an average read numbers per sample of 797,561 in Chapter 2, to an average read number per sample of 15,709,445 in the current study. This allowed for detection of Enterobacteriaceae in samples where the concentration detected by microbiological testing was below the LoD determined in Chapter 2. If we assume a linear relationship between read number and LoD, this study

would yield an LoD of $5*10^3$ CFU or PFU per sample based on the data collected in Chapter 2. This LoD is still less sensitive than microbiological methods.

The presence of Enterobacteriaceae and *Listeria* in samples recorded by NGS that were negative by traditional microbiological testing may be due to several factors. The presence of dead cells or free RNA will be detectable using molecular methods, including NGS, but would not be detected by traditional culture-based approaches. NGS will also record viable but non-culturable (VBNC) organisms. VBNC cells are living cells that are no longer able to produce colonies on rich laboratory media, and have a lower metabolic rate and different gene expression patterns to culturable bacteria (Li *et al.* 2014). In this way the VBNC state is believed to minimise energy requirements (Oliver 2010) enhancing survival of the cells in an environment of stress, such as on fresh produce or in the fresh produce environment. Many human pathogenic organisms, such as *Campylobacter*, *Salmonella* and *L. monocytogenes*, have been shown to have VBNC forms and it is notable that, for *Salmonella* and *L. monocytogenes*, these forms can be induced by chlorine stress (Magajna and Schraft 2015; Highmore *et al.* 2018). In general, VBNC bacteria are also more resistant to physical lysis and chemical stress than their culturable counterparts (Signoretto *et al.* 2000), an important observation for food associated human pathogenic organisms.

The microbiome of all samples was dominated by *Alteromonas*. This is notable as *Alteromonas* is most often associated with saltwater or saline environments (Quesada *et al.* 1983). The apparent prevalence of *Alteromonas* maybe an artefact resulting from the misassignment of reads using Kraken. Although this study subtracted the full genome of the matrix where available, for many samples it was possible only to subtract chloroplast genome from the data, and for some even this was not obtainable. Despite this a high proportion of reads were unclassified by Kraken (Table 11) which suggests matrix sequence remained within the filtered sequence reads. The similarity of the matrix genome, notably the chloroplast, to some bacterial genes can lead to misassignment of reads (see Chapter 2 Section 2 Limit of Detection and Method Comparison). The top taxa found across all samples were *Alteromonas, Pseudomonas, Leuconostoc* and *Rahnella*. There were differences in the top taxa identified in the different produce types. In the leafy green microbiome, the top taxa were *Alteromonas* and *Pseudomonas;* for the onion microbiome, the top taxa were *Alteromonas, Leuconostoc, Rahnella* and *Pseudomonas;* and for the spring onion microbiome, the top taxa were *Alteromonas, Pseudomonas*, *Leuconostoc* and *Thiobacillus*.

These findings are supported by the literature, which found the leafy green microbiome, for many produce types including rocket and lettuce, and the onion microbiome, are dominated by *Pseudomonas* (Frohling *et al.* 2018; Yurgel *et al.* 2018; Cernava *et al.* 2019). It has also been reported that the relative abundance of *Pseudomonas* increases upon refrigeration of the product (Tatsika *et al.* 2019). This observation is consistent with the high levels of *Pseudomonas* in this study where produce was refrigerated prior to analysis. *Leuconostoc* and *Rahnella* are both commonly associated with onions and have previously been described as potential onion plant pathogens (Bonasera *et al.* 2017; Asselin *et al.* 2019) but have not been described as a key member of the microbiome of leafy greens. *Leuconostoc* has also been noted to be a spoilage organism associated with food (Andreevskaya *et al.* 2018). *Thiobacillus* has previously been described in agricultural soils (Chapman 1990). The present study is the first known research to examine the spring onion microbiome.

Lefse allowed for the interrogation of the data obtained in this study to probe for members of the microbiome correlated with sample metadata. Lefse identified nine taxa (*Rahnella, Pectobacterium, Yersinia, Dickeya, Serratia, Buchnera, Lactobacillus, Pantoea* and *Enterobacter*) associated with microbiological detection of Enterobacteriaceae and *Listeria* and the detection of AMR associated genes. Eight of these taxa were part of the family Enterobacteriaceae. This is unsurprising as the microbiome of fresh produce has previously been described as containing high levels of Enterobacteriaceae (Frohling *et al.* 2018; Yurgel *et al.* 2018). The correlation between the presence of Enterobacteriaceae in the microbiology data and the presence of several Enterobacteriaceae genera in the NGS data also acts as an internal control of the statistical methods used. The correlation of Enterobacteriaceae detected by NGS with *Listeria* spp. detected by microbiological testing is a likely reason this is used as an indicator organism, but Enterobacteriaceae were also found in many samples not containing *Listeria* spp., therefore they are unlikely to give us clear information on the presence of human pathogenic organisms within the fresh produce microbiome.

The association found by Lefse between the presence of plant pathogenic bacteria (including *Pectobacterium* and *Dickeya*) and both Enterobacteriaceae and *Listeria* spp. may point to a common source or survival characteristic between these organisms or indicate that plant pathogens increase survival of human pathogens within the fresh produce microbiome. This suggests that plant pathogens are worthy of further exploration as potential indicator

organisms, with their higher titre allowing detection of these species using methods such as NGS where the LoD is too low to detect the human pathogens themselves. Interestingly, *Pectobacterium* has also been shown to enhance the survival of *Salmonella* (Wells and Butterfield 1997) and *Escherichia coli* O157:H7 (Brandl 2008) on fruit and vegetables. It has previously been suggested this increase in survival is due to the soft rot-associated release of nutrients in the local environment of the infection zone (Brandl 2008).

Lefse also found an association between presence of *Lactobacillus* and microbiological positive Enterobacteriaceae and *Listeria* spp. screens. *Lactobacillus* has long been known to produce antimicrobial compounds which inhibit growth of numerous species found within the fresh produce microbiome (Price and Lee 1970). Therefore, the presence of antimicrobial compounds produced by *Lactobacilli* may lead to a decrease in the microbial load and diversity on fresh produce, allowing for decreased competition within the microbiome, and in this way may enhance the ability of Enterobacteriaceae and *Listeria* spp. to survive. This conclusion may be supported by research finding that a less diverse soil microbiota leads to enhanced survival of bacterial human pathogens (van Elsas *et al.* 2012). The findings that *Lactobacillus* is associated with microbiological positive Enterobacteriaceae and *Listeria* spp. is also of interest as previous literature has focused on the use of *Lactobacilli* have been used as a biocontrol, or bio-preservative for fresh produce (Jamuna *et al.* 2005). The findings in the current study may suggest that the incorrect application of *Lactobacilli* may lead to enhanced survival of human pathogens. This finding is not supported by much of the literature, where *Lactobacilli* have been shown to produce antimicrobial compounds that are effective against human pathogens (Cleveland *et al.* 2001). It has also been shown that *Lactobacillus* can be used as a biocontrol agent, decreasing levels of various human pathogens (Olaimat and Holley 2012; Iglesias *et al.* 2017), including *L. monocytogenes* (Martinis and Franco 1998) and *E. coli* (Ogunade *et al.* 2016), within the food microbiome.

Leuconostoc, like *Lactobacillus*, are lactic acid bacteria that produce antimicrobial compounds (Daba *et al.* 1991). *Leuconostoc* was only found in samples with detectable levels of Enterobacteriaceae (Appendix C, i), therefore indicating a common source or survival conditions. The association between presence of *Leuconostoc* and Enterobacteriaceae microbiological positives may be for similar reasons as outlined for *Lactobacillus*. In addition, *Leuconostoc*'s role as a spoilage organism of vegetable-based foods (Vihavainen *et al.* 2008)

may lead to a greater accessibility of nutrients allowing enhanced survival of human pathogens, as previously hypothesised for the correlation between Enterobacteriaceae and plant pathogenic bacteria. Importantly, *Leuconostoc* was also found to be associated with negative microbiology results for *Listeria* species. This may be due to antimicrobial compounds produced by *Leuconostoc* which are active against *Listeria* spp. (Harding and Shaw 1990). This finding is in line with the literature where *Leuconostoc* has been found to have a bioprotective effect in fresh produce against *L. monocytogenes* (Trias *et al.* 2008a), and been shown to be a bio-preservative in some foods (Bah *et al.* 2019). This finding highlights the biological differences between bacteria associated with food and highlights the importance of further research into the total, and potentially conflicting, effects of biological control agents within the fresh produce microbiome.

From Lefse, the genera correlated with the samples negative for Enterobacteriaceae were *Pseudonocardia*, Ophiovirus and Varicovirus. Of these, Ophiovirus and Varicovirus were only present in 3 samples and therefore there is not strong evidence that these represent a true effect. *Pseudonocardia* was identified in 41 samples, with the relative abundance significantly higher in samples where Enterobacteriaceae were not detected by microbiological methods. *Pseudonocardia* is a protective mutualist within ants and is also found in soil (Holmes *et al.* 2016). *Pseudonocardia* been shown to produce anti-fungal and antibiotic compounds (Jafari *et al.* 2014). Interestingly, in the present study, there was an inverse correlation between *Pseudonocardia* and Enterobacteriacea and *Listeria* spp. consistent with the known antimicrobial action of *Pseudonocardia* and the fact that human pathogens have been shown to be susceptible to the compounds produced by *Pseudonocardia* (Jafari *et al.* 2014). These observations suggest further research on the potential afforded by *Pseudonocardia* as a biocontrol agent should be pursued.

Although Lefse is a useful methodology to identify potential correlations between the microbiome and associated metadata, the statistical methods used within the Lefse program are skewed by the presence of zeros in the microbiome data (i.e. samples where the species in question are not detected). While in the present study microbiome data was filtered to remove members of the microbiome only present in a single sample, the data could not be filtered further without biasing the statistical analysis. A larger sample size may allow for greater filtering without bias. As well as this, there is a bias when examining metadata where there are low sample numbers within one of the groups, for example in this study only having

seven samples positive for AMR associated genes of greater than 80% coverage and identity. This bias may lead to members of the microbiome being found to be significantly correlated with metadata, when this is in fact just an artefact of low sample numbers. The statistically significant outputs obtained from Lefse in this study, although significant, gave very small changes in abundance (Appendix C) and so the results need further exploration to verify their biological significance. Further work is needed to examine the genera of interest identified by Lefse for the effects these genera have on the other members of the microbiome and ascertain whether it is of biological relevance within the fresh produce microbiome.

3.5 Conclusions

The fresh produce microbiome was dominated by several key taxa, *Alteromonas, Pseudomonas*, *Leuconostoc* and *Thiobacillus.* The methodologies employed identified four AMR associated genes, *H-NS*, *CRP*, *mexF* and *mexB*, although further work could be done to examine this dataset for novel AMR associated genes. The statistical analysis of the microbiome in conjunction with metadata regarding the produce category, microbiological presence of Enterobacteriaceae and *Listeria*, and AMR gene presence yielded several genera which may affect the survival of human pathogens within the fresh produce microbiome, or show potential as an indicator organism, and could be of interest for future research. The most notable findings are the presence of *Pseudonocardia*, which correlates with the absence of Enterobacteriaceae, and the presence of Leuconostoc, which correlates with the absence of *Listeria* spp., and therefore may be utilisable as a novel method of biocontrol of human pathogens within the fresh produce microbiome. The use of *Lactobacillus* as a biocontrol agent may also lead to an increase in survival of human pathogens within the fresh produce microbiome, therefore more work is needed to ensure strains and concentrations used allow for removal of human pathogens.

Chapter 4. Phenotypic and genotypic study of *Listeria monocytogenes* isolated from vegetables, meat and clinical cases in the UK

4.1 Introduction

Listeria monocytogenes can survive and reproduce in a variety of habitats, including on food and within the food processing environment, and can also reproduce at refrigeration temperatures (Chan and Wiedmann 2009; Leong *et al.* 2014). The ability of *L. monocytogenes* to form biofilms can increase its ability to persist both in the environment and on food (Ferreira *et al.* 2014). Biofilm formation also increases an isolate's resistance to biocides thereby increasing its persistence (Ölmez and Temur 2010). Biofilm formation may also drive the acquisition of antimicrobial resistance (AMR). This may be caused by the biofilm matrix preventing antimicrobials coming into contact with cells, leading to the presence of sub-lethal concentrations of these compounds in the environment or through the close contact of cells within the biofilm increasing the potential for conjugation and the exchange of AMR associated genes (Rodriguez-Lopez *et al.* 2018). The mechanisms of action of many antibiotic resistance genes, for example efflux pumps, may also allow for survival of *L. monocytogenes* in the presence of biocides.

Isolates of antibiotic resistant *L. monocytogenes* have been found in the food supply chain in various countries throughout the world (Obaidat *et al.* 2015; Escolar *et al.* 2017; Noll *et al.* 2018; Wilson *et al.* 2018). In addition, there is growing concern over the potential of multidrug resistant bacteria within the food supply chain (Noll *et al.* 2018). The presence of multidrug resistant bacteria within the food supply chain is an issue as this will affect the treatment potential thus affecting clinical outcomes of *L. monocytogenes*. In addition, AMR associated genes may pass from foodborne bacteria into bacteria with high clinical relevance, leading to a generalised increase in morbidity and mortality due to bacterial infections.

The virulence potential of *L. monocytogenes* influences the infection potential and clinical outcomes of listeriosis. Virulence in *L. monocytogenes* is predominantly mediated by genes involved in cell to cell transfer or evasion of the host immune system (Nishibori *et al.* 1995). It has been demonstrated that plant-based molecules can inhibit the expression of virulence associated genes in *L. monocytogenes*, suggesting they are not needed for survival in the

phyllosphere, although this does not affect their expression *in vivo* (Kathariou 2002). There is limited understanding of the genes needed for survival in the phyllosphere and of whether these are linked to presence or absence of specific virulence factors, although it has previously been shown that the presence of specific genes can mediate both stress survival and virulence (Wonderling *et al.* 2004).

As previously discussed, whole genome sequencing (WGS) is increasingly being utilised in the context of food microbiology and pathogen identification (U.S. Food and Drug Administration 2019) and has been used successfully to track outbreaks of *Listeria monocytogenes* within the food supply chain (Public Health England 2019). This services a significant volume of data that can be utilised to screen for phylogenetic relationships between isolates, and potential phenotypic qualities, for example resistance or virulence associated genes (Smith *et al.* 2019). Genome wide association studies (GWAS) can also beundertaken to probe the presence of novel genes associated with a specific phenotype or piece of metadata, for example the origin of the isolate.

The aims of this study were to:

(i) Qualify, through phenotypic testing, the antibiotic resistance profile and biofilm-forming capability of *L. monocytogenes* isolated from fresh produce.
(ii) Compare WGS data of *L. monocytogenes* isolated from the fresh produce supply chain with similar from UK-related meat and clinical isolates from the NCBI Sequence Read Archive to identify the multi-locus sequence typing (MLST) types most frequently associated with each source, their phylogenetic relationships, and the rates of incidence of AMR and virulence genes.
(iii) Employ GWAS on WGS data and metadata on the category of origin (fresh produce, meat or clinical), to identify genes associated with the origin of the sample.

4.2 Methods

4.2.1 Phenotypic screening of L. monocytogenes isolated from fresh produce

4.2.1.1 Selection of Bacterial Strains

Listeria monocytogenes strains isolated from fresh produce (see Appendix D for information on source of isolates) were supplied by Edinburgh Napier University (isolates veg1-14 appendix D) and a UK-based fruit and vegetable manufacturer (isolates veg15-48 appendix D). The strains were isolated during routine food testing between May 2016 and November

2018 following ISO 11290-2: 2017 and were identified using biochemical testing or MALDI-TOF. Strains were received streaked on nutrient agar slopes. The isolates were then sub-cultured at Fera onto Chromogenic *Listeria* Agar (OCLA) (Thermo Scientific, Oxoid, Basingstoke, UK) and grown for 48 h at 37 °C enabling preliminary confirmation as *L. monocytogenes* and assessment of purity. Individual colonies were then recovered into Protect tubes (Technical Service Consultants) containing liquid cryoprotectant buffer plus plastic beads (aiding viability of frozen cultures) prior to storage at -80°C. *Streptococcus pneumonia* strain NCTC 12977 was obtained from the PHE culture collection to use as a positive control. The strain was plated onto Columbia Blood Agar (CBA) and grown overnight at 37 °C to obtain an active culture and assess purity, then a 1 µl loopful of culture placed into a Protect tube and stored at -80°C.

4.2.1.2 Phenotypic Screens

Isolates of *L. monocytogenes*, plus *Streptococcus pneumoniae* per batch tested, were removed from -80 °C storage and a 1 µl loopful of cryoprotectant containing the culture plated on Columbia Blood Agar (CBA) and grown overnight at 37 °C. Bacteria were taken from this plate and resuspended in phosphate buffered saline (PBS) to McFarlan 0.5 standard; equivalent to an absorbance of 0.08 at OD650. This suspension was then used for subsequent phenotypic screens as in Figure 11 following EUCAST guidelines for disk diffusion (The European Committee on Antimicrobial Susceptibility Testing 2017).

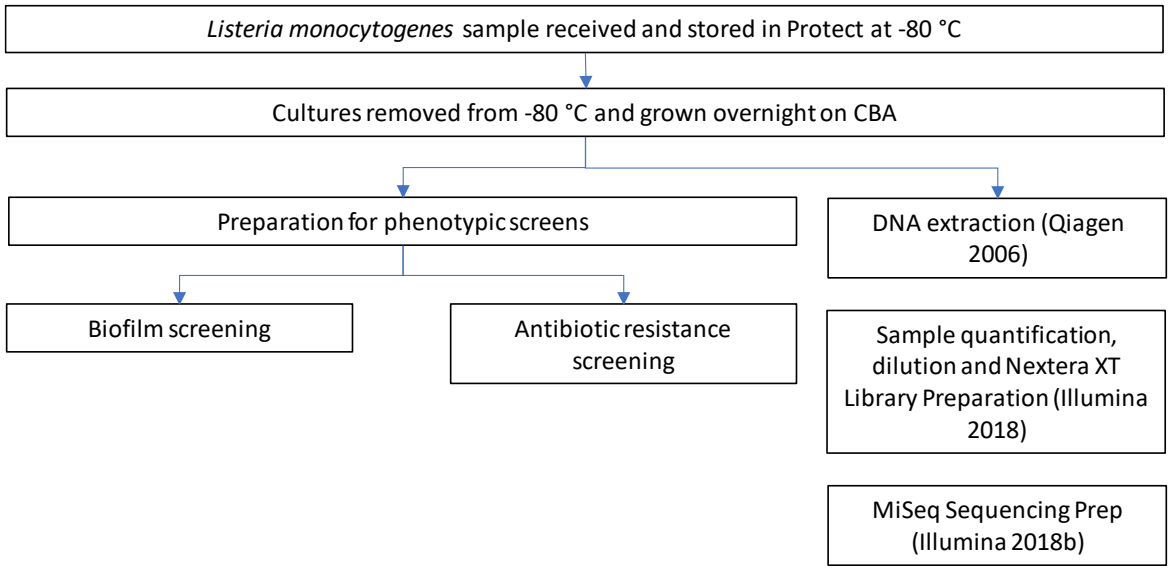


Figure 11. Flow chart showing the processing of *Listeria monocytogenes* isolates upon receipt through phenotypic and whole genome sequencing methods.

Briefly, to test for antimicrobial resistance; resuspended isolates were inoculated onto Mueller Hinton Media with Horse Blood (MH-F)(Thermo Scientific, Oxoid), following EUCAST guidelines. A sterile swab was dipped into the prepared culture dilution, and swabbed onto the plate in three directions to spread inoculum evenly over entire plate. Antibiotic disks of Penicillin (P), Ampicillin (AMP), Meropenem (MEM), Erythromycin (E), Trimethoprim-sulfamethoxazole (SXT) and a blank were taken from their packaging using sterile forceps and place directly onto plate. Plates were incubated at 37 °C for 18 h in stacks no larger than 6 plates. After 18 h, electronic callipers were used to read the size of the zone of clearance around the antibiotic disk. Resistance and non-resistance breakpoints were categorised using the EUCAST breakpoints for *Listeria monocytogenes.*

To test the ability of the isolates to form biofilms, 1 ml of liquid media, brain heart infusion (BHI) or nutrient broth (NB) was added to appropriate wells of a 24 well flat-bottomed ELISA plate. To this, 100 µl of absorbance 0.08 at OD650 resuspended culture was added, with each culture performed in triplicate for each medium, and one well left as a blank per row. This was incubated at 35 ^{o}C for 24 h, then the liquid media removed, the plate washed three times with 1 ml of PBS and stained with 200 µl 0.1% crystal violet (Kadam *et al.* 2013) at room temperature for 45 minutes. After this the excess crystal violet was removed and the plate washed three times with 1 ml of sterile distilled water (SDW). To de-stain the biofilm, 200 µl of 95-98% ethanol was added to each well and incubated for 30 minutes at room temperature, then a 100 µl aliquot transferred to a 96 well microplate and read in a plate reader at 595 nm. Absorbance data were plotted to assess the ability of the isolates to form biofilms (Appendix E).

All phenotypic screens were performed in triplicate with a control plate of *S. pneumoniae* run for each batch of samples tested. Standard deviation, minimum values and maximum values for each phenotypic test were calculated using Microsoft Excel.

4.2.2 Whole Genome Analysis

4.2.2.1 Preparation of L. monocytogenes isolated from fresh produce

For WGS; isolates of *L. monocytogenes* from fresh produce were taken from Protect, plated on Columbian Blood Agar (CBA) and grown overnight at 37 ^{o}C. Two culture collection strains (NCTC 11994 and NCTC 5214) were also plated and grown overnight at 37 ^{o}C. DNA extractions were performed on a single colony using the Qiagen Blood and Tissue kit following manufacturer's instructions for Gram-positive bacteria (Qiagen 2006). Lysis buffer

was made up by mixing 0.1g Lysozyme, 60 µl TritonX, 100 µl Tris-EDTA and 5 ml MBGW. A single colony of *L. monocytogenes* was taken from the CBA plate and resuspended in 180 µl lysis buffer, then incubated for 50 min at 37 °C. To this, 25 µl proteinase K and 200 µl buffer AL were added and the mixture incubated for 30 mins at 56 °C, followed by a further 15 mins at 95 °C. After this, 200 µl of 100% ethanol was added and the mixture loaded onto a DNeasy spin column. Manufacturer's instructions (Qiagen 2006) were followed and the DNA eluted in 50 µl, followed by a further 50 µl of buffer EB.

The samples were quantified using the Qubit® DNA HS Assay Kit following the manufacturer's instructions and diluted to 0.8 ng in 5 µl. Samples were prepared for WGS using the Nextera XT Library Preparation Kit following the manufacturer's instructions (Illumina 2018c), with an input of 0.8 ng DNA. To the diluted DNA, 10 µl buffer TD and 5 µl ATM per sample were added, the plate sealed and placed on a thermal cycler for: 5 mins at 55 °C , held at 10 °C. Upon reaching 10 °C, 5 µl NT buffer was added to each well and mixed gently by pipetting up and down 10 times. The mixture was incubated at RT for 5 mins before addition of 15 µl NPM to each well, plus 5 µl each of unique forward and reverse index primers. The plate was sealed and place on thermal cycler for 72 °C for 3 mins, 95 °C for 30 s, followed by 13 cycles of 95 °C for 10 s, 55 °C for 30 s and 72 °C for 30 s, before a final elongation step at 72 °C for 5 mins. Samples were cleaned via the addition of 30 µl AMPure XP beads, incubated at RT for 5 mins, placed on magnetic stand and supernatant removed, before being washed twice with 80% ethanol, and resuspended in 38 µl MBGW.

Samples were then quantified using the Qubit® DNA HS Assay Kit following the manufacturer's instructions and pooled to equimolar concentrations. Pool quality was checked using the Agilent 2200 TapeStation system with High Sensitivity D1000 reagents (Agilent Technologies) following the manufacturer's instructions. The pool was denatured, combined with 5% PhiX and diluted to 10 pM then run on a single MiSeq flow cell using the V3 reagents kit (Illumina).

4.2.2.2 Quality control

Extraction blanks were undertaken as part of the RNA extraction. Process blanks, MBGW put through the same processing as samples, and indexing blanks, MBGW added at indexing PCR stage, were undertaken for each sequencing method. All were examined using the tapestation and Qubit for quality purposes. Blanks were run on the sequencer as a separately indexed sample. All samples, including blanks, were examined for read number

and those with low quality reads, or less than 500 reads were filtered out of the analysis. Any blanks remaining were run through the bioinformatic analysis separately to samples and the top taxa compared manually to those in experimental samples to rule out cross contamination.

The PhiX internal standard was spiked into the final pool and run on the sequencer. The PhiX standard was mapped to the PhiX genome on the MiSeq as part of the standard Illumina workflow to allow for the assessment of the quality of the MiSeq run, in addition to metrics on cluster density and read numbers. This was compared to the average statistics on these metrics for the specific run type (amplicon or metagenomics) runs on the MiSeq at Fera to ensure the quality of the run was of the standard usually obtained.

4.2.2.3 Bioinformatic Analysis

WGS data from 46 clinical samples and 34 meat isolated samples were downloaded from the NCBI Sequence Read Archive. Details of origin and NCBI reference number are outlined in Appendix D.

Initially, to examine the relatedness of the strains (as outlined in Figure 12), the samples were trimmed using Sickle v1.33 (Joshi and Fass 2011) to remove sequence of quality less than Q20 (1 in 100 probability of incorrect base call) and length less than 100 base pairs. The genomes were then assembled using Spades v3.10.1 (Nurk *et al.* 2013). Core genome MLST (cgMLST) assignment was undertaken using chewBBACA (Silva *et al.* 2018) and then visualised in PHYLOViZ (Francisco *et al.* 2012) in conjunction with metadata related to the origin of the sample. Average nucleotide identity was undertaken using PyANI (Pritchard *et al.* 2016).

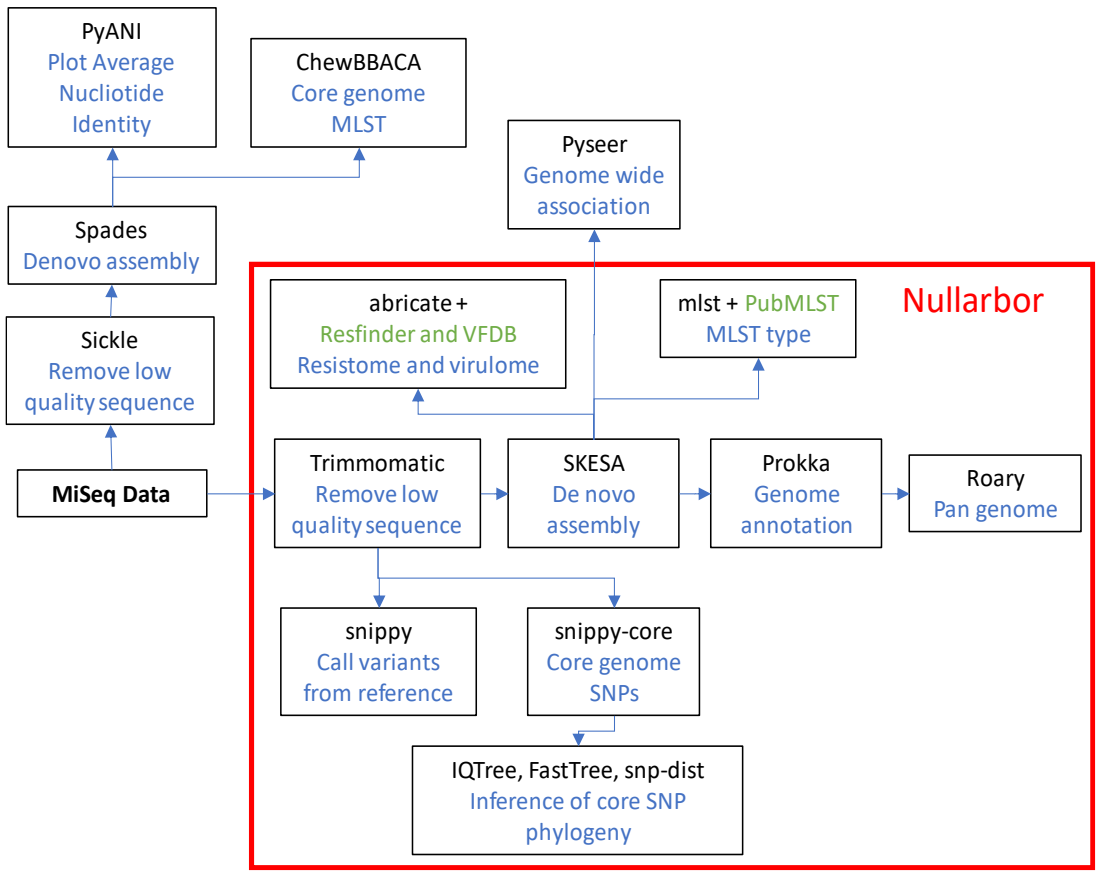


Figure 12. Flow chart of bioinformatic methods used to analyse whole genome sequencing data of *Listeria monocytogenes*.

Analysis of MLST type, assignment of the resistome and virulome, and the phylogeny of isolates were examined using Nullarbor version 2.0.20181010 (Seemann *et al.* 2018), a pipeline designed to generate public health microbiology reports from WGS isolates, with *Listeria monocytogenes* strain NCTC10357 (NZ_LT906436.1) as the reference genome (Figure 12). The Nullarbor pipeline briefly includes trimming the sequence using Trimmomatic, use

of trimmed data to call variants using snippy and examine core genome single nucleotide polymorphisms (SNP) using snippy-core, which then allowed for the creation of an SNP phylogeny tree using IQTree. The trimmed sequences underwent *de novo* assembly using SKESA. Assembled reads were used to assign MLST type using the software mlst with the PubMLST database, to infer the resistome and virulome using abricate plus Resfinder and VFDB databases and were annotated using Prokka. The annotated reads were used to identify the pan genome of the isolates using Roary. *De novo* assembly outputs from Nullarbor were also used as inputs for Pyseer (Lees *et al.* 2018), to compare genomes and obtain genome wide association (GWAS) data for each origin.

4.3 Results

4.3.1 Phenotypic screening

None of the isolates tested appeared to exhibit resistance to any of the antibiotics used when assessed against EUCAST breakpoints (The European Committee on Antimicrobial Susceptibility Testing 2018). There was variability in the extent of clearance zones observed (Appendix F), but all were above the breakpoints (Figure 13). However, the minimum zone of clearance for both AMP and SXT was within 0.25 mm of the breakpoint. There were no statistically significant differences between the resistance profiles of the two lineages (lineage identified by WGS).

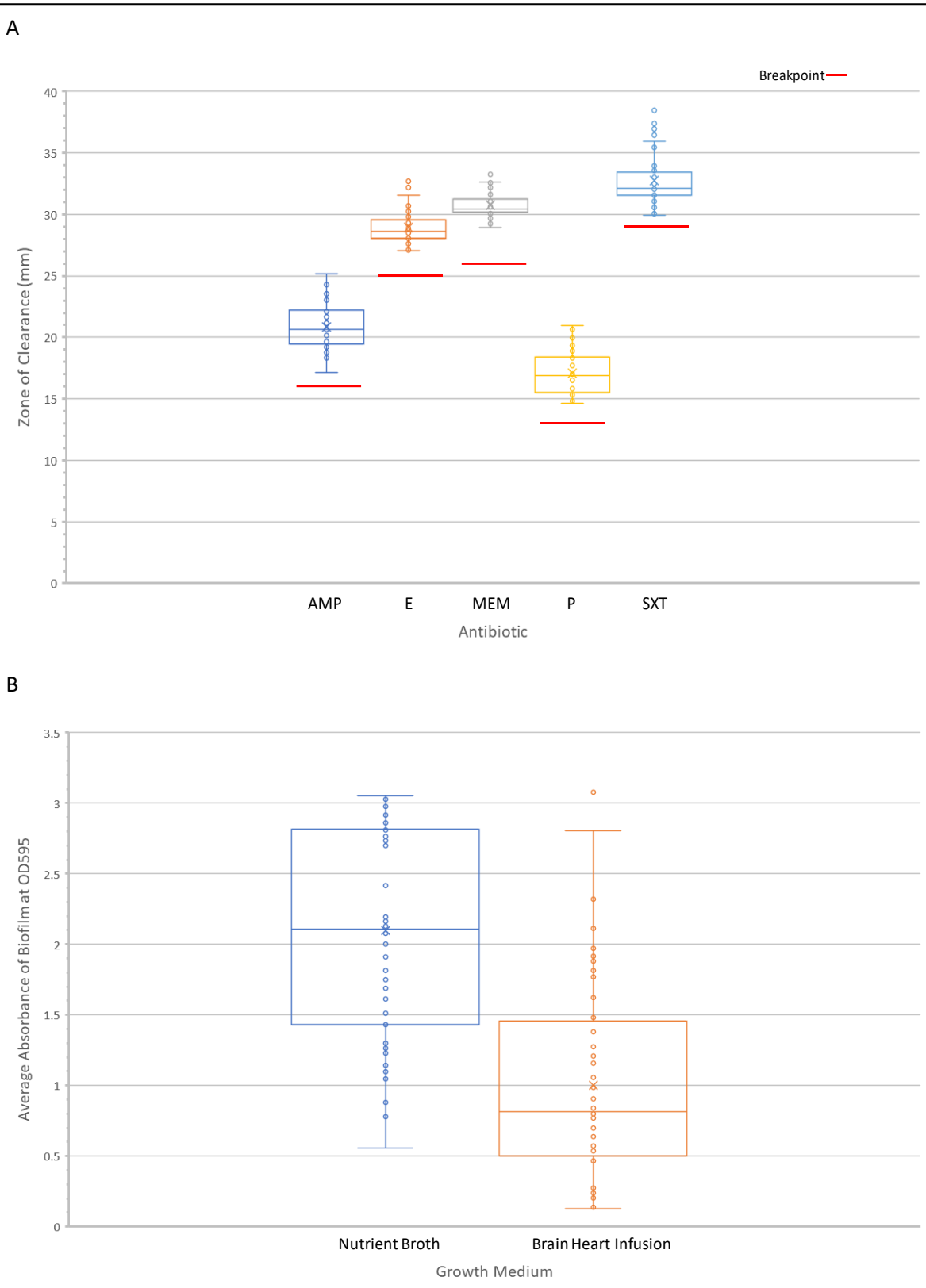


Figure 13. Phenotypic screening for: A) antibiotic resistance and B) biofilm formation capability. AMP = Ampicillin, E = Erythromycin, MEM = Meropenem, P = Benzylpenicillin, SXT = Trimethoprim-sulfamethoxazole. Antibiotic resistance is measured in mm, biofilm formation is measured in OD at 595 nm.

All isolates tested were able to form biofilms in both NB and BHI. The biofilm formation screens showed a highly significant difference in biofilm formation capability in NB and BHI (Students Paired T-Test $p \leq 0.0001$); *L. monocytogenes* forming biofilm better in NB than in BHI (Figure 13), which is consistent with the observation that stress conditions favour biofilm formation. The maximum absorbance reading was similar in both media, but the minimum absorbance was lower for BHI than NB (Figure 13). The ability to form biofilms in NB varied significantly (Students T-Test $p \leq 0.0001$) between isolates from lineage 1 and lineage 2.

4.3.2 WGS

4.3.2.1 Quality Assessment

Run quality metrics are found in Table 13. All runs had a high level of reads passing filter and assigned to indexes, indicative of a good quality run. Negative controls were examined, and all had read numbers of less than 500, therefore were filtered out at QC stage and no subsequent analysis was performed on them.

Table 13. MiSeq run metrics for each run associated with data from whole genome sequencing of *Listeria monocytogenes* samples

Cluster Density	Reads Passing Filter	% Clusters Passing Filter	% PhiX Loaded into Library	Concen-tration of library loaded	% of Reads Aligned to PhiX	Error Rate	%Q30	% Identified	% Assigned to Index
1189	25395640	90.83	5	10	7.09	3.43	78.84	90.8	2539655
1054	25530000	93.24	10	10	20.18	3.59	76.6	72.23	2553072

4.3.2.2 Identification and relatedness

The WGS data showed all samples belonged to either lineage I (82 isolates) or II (46 isolates), except culture collection sample NCTC 5214, which was lineage III (Figure 14). The proportion of isolates from each category of origin varied between the two lineages. Of the isolates in lineage I, 42% were from fresh produce, 20% were from meat and 35% were from clinical cases, compared to lineage II where only 28% of the isolates were from fresh produce, 37% from meat and, as with lineage I, 35% were from clinical cases. Moreover, there was greater genetic variation in isolates from lineage II than lineage I as shown using average nucleotide identity plots (Figure 15). There were 35 MLST sequence types (ST) (Figure 14), with ST6 constituting the most frequently found ST (33 isolates), then ST 9 (14

isolates), and ST1 (13 isolates). The cgMLST was found to give greater differentiation of strains than conventional MLST. When the relatedness, as assigned by cgMLST, was visualised coloured by MLST type (Figure 16) the MLST types cluster within the cgMLST phylogeny. None of the methods tested differentiated the origin of samples, and none of the STs clustered to a single origin (where n>2), except for ST3, which was assigned to 4 isolates from meat, but none of vegetable or clinical origin, and ST 7, which was assigned to 4 isolates from clinical origin, but none of meat or vegetable origin. In addition, samples did not phylogenetically cluster based on origin (Figure 14).

A total of 7185 coding sequences (CDS) were found in this dataset by Nullarbor, with this ranging from 2785-3073 CDS within each isolate. The core genome (genes found in 99-100% of isolates) consisted of 2103 genes, a further 132 were found in 95-99% of strains, 1245 found in 15-95%, and 3701 found in 0-15% of strains.

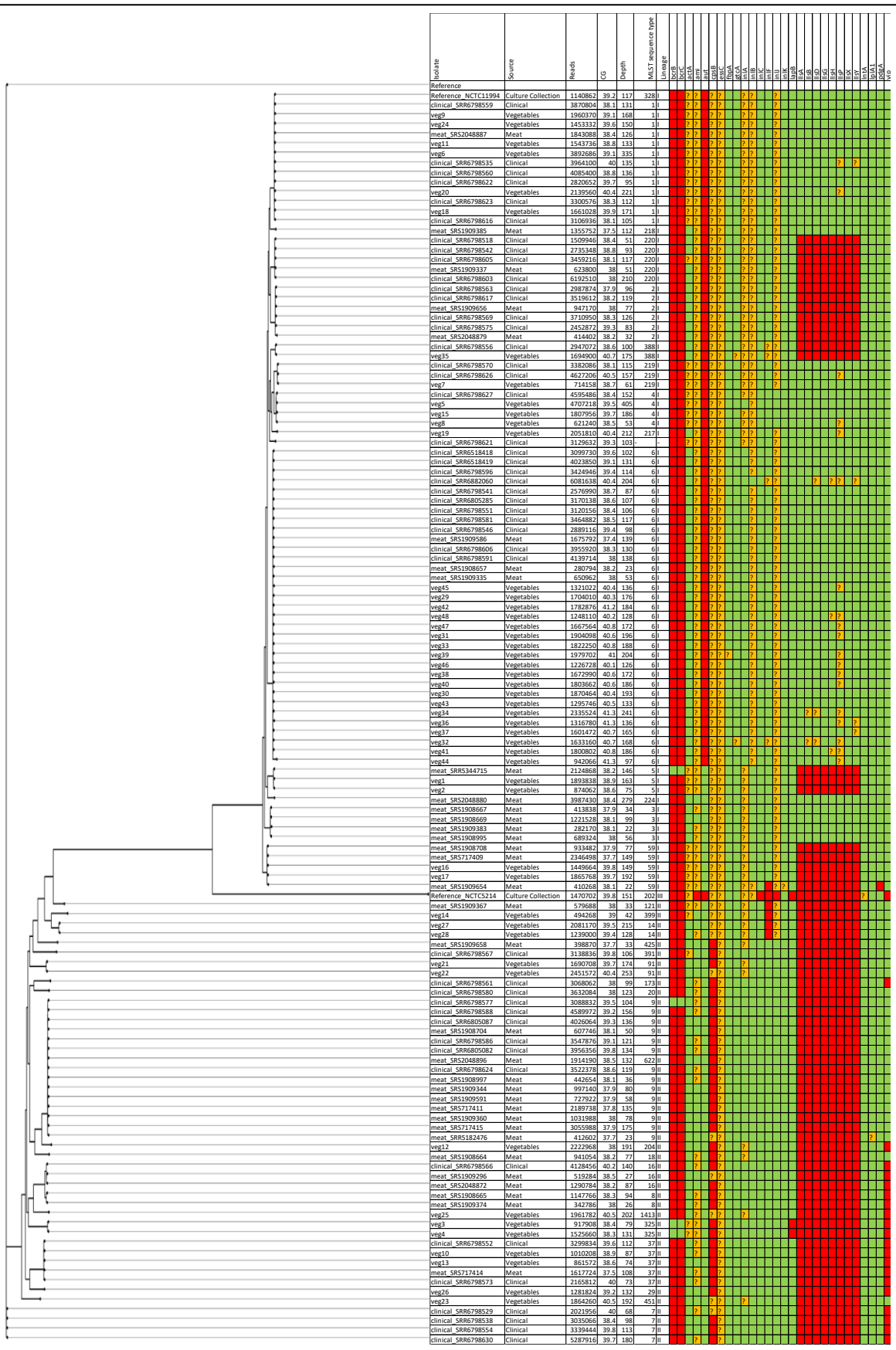


Figure 14. Phylogenetic tree produced by core SNP analysis via Nullarbor. Number of reads associated with the sample, CG content, depth, MLST sequence type and lineage, and the distribution of genetic elements associated with antimicrobial resistance and virulence. Shading on the right of the tree indicate the presence (green), absence (red), or potential presence (yellow) of genes.

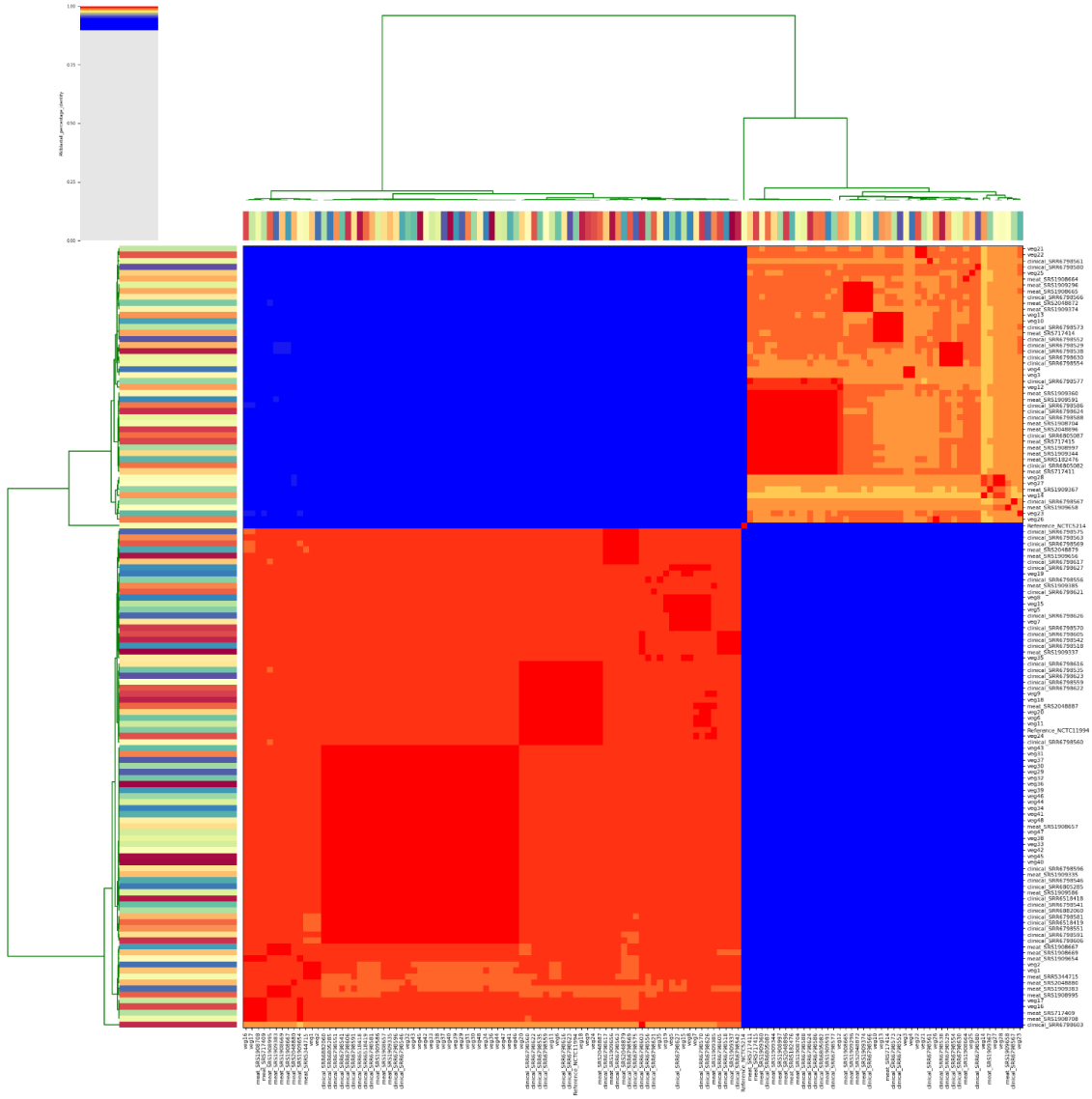


Figure 15. PyANI output showing the genetic relatedness of *Listeria monocytogenes* isolates tested, with red being most closely related (>98% ANI).

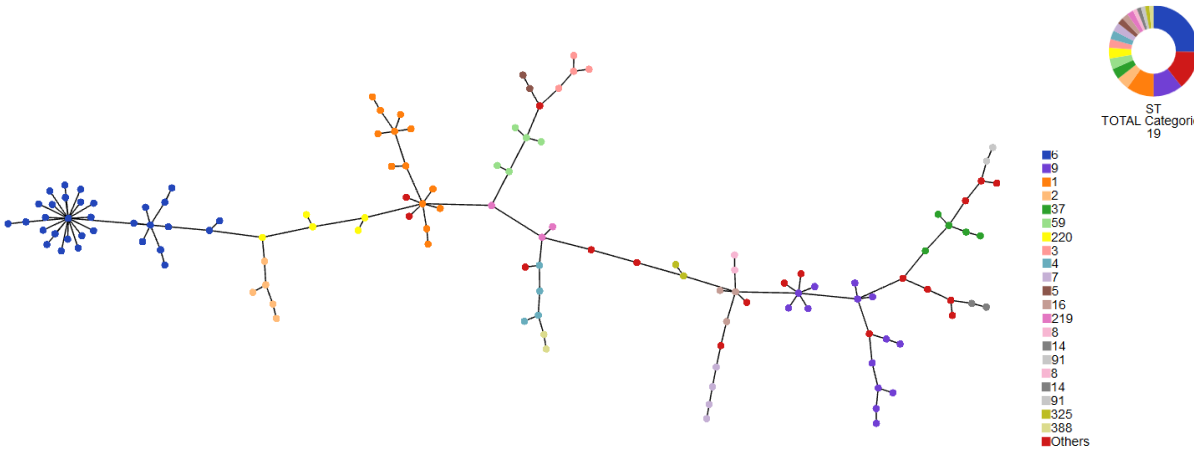


Figure 16. cgMLST of *L. monocytogenes* isolates assigned using chewBBACA and visualised in PHYLOViZ, coloured based on MLST type given by Nullarbor.

4.3.2.3 Phenotypic Inference

The antibiotic resistance genes *fosX* (conferring resistance to Fosfomycin) and *lin* (conferring resistance to Lincosamide) were identified in all isolates. The genes *bcrB* and *bcrC* (both conferring resistance to Bacitracin) were found together in 4 isolates (Figure 14). There were 42 genes found in the virulome. Of the genes found, 17 were ubiquitous amongst the isolates tested (*bsh, clpC, clpE, clpP, essC, hly, hpt, iap/cwhA, lap, lpeA, lspA, mpl, oatA, plcA, plcB, prfA, prsA2*), and two (*cpsB* and *essC*) exhibited 16% coverage and 78% identity to the corresponding assigned gene.

Genes in *L. monocytogenes* pathogenicity island III (*llsA, llsB, llsD, llsG, llsH, llsP, llsX, llsY*), encoding Listeriolysin S and associated proteins, were not found in isolates belonging to lineage II or III. In lineage I there are clusters of isolates containing (n= 61) or missing (n= 21) the full pathogenicity island. None of the antibiotic resistance or virulence genes found were exclusively found in isolates belonging to a single category of origin (Figure 14).

GWAS highlighted 218 genes significantly (p≤ 0.00000001) associated with isolation from vegetables (Appendix G, i), with the top results being *int* (an integrase from bacteriophage A118); *cds2325* (encoding a hypothetical protein of unknown function); and *cds2326* (an anti-repressor with limited information known about its function). Many of the top results identified as hypothetical proteins with little or no information known about their function. GWAS identified 11 genes significantly (p≤ 0.00000001) associated with *L. monocytogenes* isolated from clinical samples (Appendix G, ii), with the top results being *cds460* (a

hypothetical protein) and *cds456*, (a transcriptional regulator). No genes were identified as significantly associated with *L. monocytogenes* isolated from meat or variation in the phenotypic screens undertaken in the laboratory including those for AMR and biofilm-producing capability.

4.4 Discussion

All *L. monocytogenes* isolates exhibited the ability to form biofilms in both NB and BHI. Consistent with previous findings, strains of *L. monocytogenes* lineage II showed a stronger capability to form biofilms than those of lineage I (Djordjevic *et al.* 2002; Borucki *et al.* 2003). However, biofilm formation was more evident in NB than BHI. This suggests that the ability of *L. monocytogenes* to form biofilms is linked to stress, as NB is a nutrient-poor, less suitable media for the culture of *L. monocytogenes* than BHI (Jones and D'Orazio 2013). Similar findings were reported by Kadam and co-workers (2013) who also noted that, as in the present study, biofilm screens performed at 37 °C are unlikely to reflect the biofilm-forming capability of *L. monocytogenes* in the food processing environment where temperatures are commonly much lower. Previous research by Bonsaglia *et al.* (2014) has shown that, although there are differences in biofilm formation dependant on temperature, these are down to the speed of growth not *L. monocytogenes*' ability to form biofilms, therefore given a longer timescale at the temperatures found in the food processing environment we would see similar results. The ability of *L. monocytogenes* to form biofilms may also be affected by other microbes in the environment (Carpentier and Chassaing 2004). Likewise, the development of multispecies biofilms may influence *L. monocytogenes'* resistance to biocides thus influencing its survival in the food environment (Bridier *et al.* 2015).

Low levels of AMR were observed in *L. monocytogenes* isolated from UK fresh produce using both molecular and phenotypic methodologies, and low levels of AMR were also evident in meat and clinical isolates assayed using molecular methods. There was no resistance observed using phenotypic testing to the clinically relevant antibiotics P, AMP, MEM, E and SXT. Only four genes were identified using WGS that are associated with AMR. Of these *fosX* and *lin* were ubiquitous in the 130 isolates tested in this study. Most strains of *L. monocytogenes* have previously been described as exhibiting innate resistance to fosfomycin and lincomycin (Aureli *et al.* 2003; Olaimat *et al.* 2018). Four isolates were found to also contain the genes *bcrB* and *bcrC*, two of vegetable origin, one of meat, and one of clinical

origin. These genes encode for resistance to bacitracin, an antimicrobial not clinically relevant for treatment of *L. monocytogenes* and therefore not tested as part of the phenotypic screens of fresh produce associated isolates. Bacitracin is an antibiotic, naturally produced by members of the Bacillus subtilis group (Johnson *et al.* 1945), that is predominantly utilised as a topical treatment for skin infections (Spann *et al.* 2004). Recorded resistance to this antibiotic is uncommon but has been previously recorded in staphylococci (Spann *et al.* 2004).

Similarly low levels of AMR have previously been reported using WGS to explore *L. monocytogenes* from European ready to eat foods and human clinical cases (Painset *et al.* 2019); and other studies reported similar finding for isolates of Italian, Polish and Australian origin (Conter *et al.* 2009; Maćkiw *et al.* 2016; Wilson *et al.* 2018). However, isolates from Iran, Germany and Spain have been reported to exhibit much higher levels of AMR (Kalani *et al.* 2015; Escolar *et al.* 2017). For example, Noll and co-workers (2018) reported significant multidrug resistance in 56% of *L. monocytogenes* isolates recovered from food samples in Germany, predominantly milk, with resistance observed to a wide spectrum of antibiotics, including several traditionally used to treat listeriosis. The regional variation in the detection of AMR genes in *L. monocytogenes* may be due to the difference in the methodologies and origin of samples between studies, with studies testing different foods and numbers of isolates. This highlights the importance of screening human pathogenic organisms for the emergence of antimicrobial resistance, something that could be achieved with the routine use of WGS.

The *L. monocytogenes* genome is highly conserved; the core genome of the isolates in this study comprised 2,235 genes. The number of CDS found in this study ranged from 2785-3073, which is consistent with other studies (Fox *et al.* 2016). This equates to 73%-80% of CDS identified in each isolate being part of the core genome. Phylogenetic analysis showed a clear separation between *L. monocytogenes* lineages and clustering of MLST STs, with cgMLST delivering greater differentiation of strains than conventional MLST (Chen *et al.* 2016). cgMLST also delivers the added advantage of enabling comparison of data between laboratories, making it a powerful (and commercial) tool for surveillance, source tracking and analysis of outbreaks. Previous work reflected the finding in this study that isolates from food and human cases belong to lineage I and II, although in contradiction to the presented

work, studies also show lineage I is made up of a greater proportion of clinical than food isolates (Painset *et al.* 2019).

Many of the virulence factors found in this study were ubiquitous to all 130 isolates examined. For those that were not ubiquitous, the presence or absence of the virulence genes were predominantly clustered based on the phylogeny of the isolates, notably for the genes found in pathogenicity island III. This reflects the limited amount of recombination reported in *L. monocytogenes* (den Bakker *et al.* 2010). There were no virulence factors found where presence or absence was associated with a single category of origin. There was no evidence of attenuation of virulence genes in food isolates compared to clinical isolates as has previously been theorised within the literature (Kathariou 2002). Further work is needed to examine samples with partial mapping to virulence genes to identify truncation and non-sense mutations, as these have previously been shown to be associated with changes in virulence (Maury *et al.* 2016; Maury *et al.* 2017). At present all strains of *L. monocytogenes* are treated equally in a regulatory capacity, but an increased knowledge of virulence patterns may allow for a greater understanding of the survival of fresh produce associated strains and their potential of to cause disease.

GWAS was able to identify a greater range of CDS associated with isolates of vegetable origin (218) compared with isolates of clinical (11) or meat (0) origin. The variation in the number of CDS significantly associated with each category of origin may reflect the genetic diversity required to survive in the produce environment or may be due to the higher number of isolates from vegetables in lineage I, which is genetically more variable. GWAS found Internalin associated with clinical case isolates, which is unsurprising as it is one of the most well described virulence genes of *L. monocytogenes* (Gaillard *et al.* 1991). Many of the other CDS identified were for hypothetical proteins, this finding reflects findings by Pirone-Davies and co-workers (2018) and highlights the need for further research into non-clinical isolates, which may house novel genes of interest. This may allow for the development of new measures to reduce the persistence of *L. monocytogenes* in/on food or within the food environment. Many of the proteins associated with either vegetable or clinical origin were hypothetical proteins, again consistent with the work of Pirone-Davies and co-workers (2018) and highlights the need for further research into non-clinical isolates, which may house novel genes of interest. GWAS was unable to identify any CDS linked to the phenotypic variations seen in the phenotypic screens performed in this study. This is possibly

due to the low levels of variability in phenotype between samples, and the low sample numbers screened (n = 48). Greater sample numbers may have yielded more information.

4.5 Conclusions

L. monocytogenes isolates from the UK food chain are genetically closely related to clinical case isolates and contain many well documented virulence genes. WGS identified four genes contributing to AMR and enabled screening of resistance to a greater number of antimicrobial compounds than would have been possible using phenotypic screens. The use of WGS, notably through cgMLST typing, facilitated greater differentiation of strains than feasible *via* conventional MLST typing, demonstrating the power of this approach and highlighting a need to implement WGS as a precautionary (rather than a reactionary) approach in routine food microbiological analyses. Deployment of WGS approaches in food testing could lead to improved source tracking, antibiotic resistance detection and the identification of potentially novel methods for biocontrol and prevention of disease.

Chapter 5. General Discussion

Throughout this thesis next generation sequencing (NGS) has been utilised, in conjunction with traditional microbiological methodologies, to detect and characterise foodborne pathogens, examine the microbiome of fresh produce and examine potential influences on the survival and transmission of human pathogens within the fresh produce supply-chain. This was achieved using both whole genome sequencing (WGS) and metatranscriptome approaches in conjunction with a variety of bioinformatic methodologies.

This project had several aims;

(i) Develop laboratory and data analysis protocols that enable the identification of the microbiome of fresh produce and identify the limits of detection of these methods for human pathogens.

In chapter 2 of this thesis, methods for the enrichment of nucleic acids, sequencing preparation and bioinformatics methodologies were compared in the context of detection of foodborne pathogens. It was found that the most sensitive limit of detection using the adopted protocols was 10^4 CFU of *Salmonella* and 10^5 PFU of MS2 per sample, a level achievable utilising enrichment via ribosomal depletion, sequence preparation for MiSeq analysis using the ScriptSeq kit, and bioinformatic analysis using Kraken in conjunction with mini-database. Additionally, a comparison of sequencing platforms was undertaken to assess the effect of sequencing platform, and differences in read length achieved, on the LoD achievable. For Illumina platforms, the HiSeq, MiSeq and NovaSeq, the LoD did not change despite differences in read length, when the read depth was kept constant. This study is of importance as it ascertains the limits of detection of current sequencing and bioinformatic techniques, where previous studies are now outdated due to their focus on technologies no longer in use, such as the 454-pyrosequencer (Frey *et al.* 2014; de Boer *et al.* 2015). Moreover, many previous studies have undertaken the examination of methods at points in the sequencing chain (Caporaso *et al.* 2012; Fouhy *et al.* 2016; Allali *et al.* 2017), with none having examined the effects of all stages of the sequencing process simultaneously, as achieved in this study. Methodologies have recently been described for the use of metagenomics for pathogen detection (Wylezich *et al.* 2018), but until the current work there were no studies examining the utilisation of metatranscriptomics for the detection of human pathogens in fresh produce, or comparing multiple methods. This can be viewed as

the first step in the validation of the potential use of NGS in food microbiology, a crucial step for standardisation and comparison between laboratories. In order to utilise NGS as a potential detection tool, standardised and validated methods are required (Endrullat *et al.* 2016; Haynes *et al.* 2019). This is important since a third of all reported foodborne contamination issues are presently of unknown causes and may therefore be due to pathogens not routinely tested for, all of which could be screened for using non-targeted approaches such as NGS (European Food Safety Authority 2018).

(ii) Analyse the fresh produce microbiome from samples obtained from the food supply chain and examine for correlations with microbiological data.

In chapter 3 of this thesis, samples were obtained from within the fresh produce supply chain analysed using the most appropriate methods ascertained in chapter 2 of this thesis, to examine the microbiome. Samples sequenced were also tested by a collaborator, using standard microbiological screening, to assess the presence and absence of Enterobacteriaceae and *Listeria* species. These data were combined and used to assess the taxa within the microbiome that were positively or negatively correlated with presence of Enterobacteriaceae and *Listeria* spp. to highlight elements which show potential to be indicators or biocontrol agents.

Nine taxa were found that were positively correlated with the presence of Enterobacteriaceae and *Listeria* species, several of which have been linked in the literature with the increased survival of human pathogens in the fresh produce microbiome (Wells and Butterfield 1997). This may be due to common factors allowing the survival of these organisms, or more likely due to the direct impact of these taxa, probably on nutrient availability via their effect on the competitive speciation of the microbiome and/or the production of antimicrobial compounds (Daba *et al.* 1991; Cleveland *et al.* 2001).

Only one taxon, *Pseudonocardia,* was found to correlate with the absence of Enterobacteriaceae. This is a bacteria which produces antimicrobial compounds, and previous research has shown human pathogens are susceptible to these compounds (Jafari *et al.* 2014). Further work is recommended to elucidate the potential of this organism as a biocontrol agent, including the examination of its effects on the microbiome, specifically human pathogens.

The microbiome sequence information collected in chapter 3 was compared to data on presence of Enterobacteriaceae and *Listeria* spp. by microbiological methods to examine the

concordance of NGS with routine microbiological methods. Of the samples tested as positive using microbiological methods, four of 31 positive for Enterobacteriaceae and all those positive for *Listeria* spp. were negative by NGS screening. This may reflect the relative insensitivity of NGS as shown in chapter 2 of this thesis. In addition, many samples testing negative by microbiological testing were found to contain Enterobacteriaceae and *Listeria* spp. by NGS approaches. These results highlight a key issue with the use of NGS approaches as a routine microbiological screening tool, the lack of consensus between NGS results and microbiological results, examined in more detail later in this discussion. Previous studies have shown that the microbiome predicted by NGS methods is influenced by the study design (Jones *et al.* 2015) and although studies have been undertaken to examine best practice in microbiome studies (Pollock *et al.* 2018), little information is available to compare the known microbiological content of foods, feeds and drinks to those reported using NGS methods.

(iii) Use phenotypic and genotypic methods to characterise the resistome, virulome, and biofilm forming ability and assess the phylogenetic identity and gene content of these isolates compared to 80 isolates of meat and clinical origin to identify signatures of fresh produce contaminating *L. monocytogenes.*

In chapter 4 of this thesis, isolates of *L. monocytogenes* isolated from the fresh produce supply chain were screened using microbiological methods to highlight phenotypes of interest, including those expressing antibiotic resistance and biofilm formation. There were broad differences between isolates of *L. monocytogenes* in their capability to form biofilms, and all were stronger biofilm formers in NB than BHI, consistent with previous findings (Borucki *et al.* 2003). No AMR phenotypes were identified in 50 isolates of *L. monocytogenes* from the fresh produce supply chain. This study represents the most extensive AMR screening yet conducted on *L. monocytogenes* from the UK fresh produce supply chain. These isolates of *L. monocytogenes* obtained from the fresh produce supply chain additionally underwent WGS. The resulting data were combined with data from the NCBI database of whole genome sequence data of *L. monocytogenes* from isolates from meat and clinical cases. Examination by WGS was able to give greater differentiation of *L. monocytogenes* strains than MLST typing, as has previously been reported in the literature (Henri *et al.* 2017). Additionally, these data were screened for genes associated with phenotypes of interest, including virulence. This work built on similar studies examining

phenotypic and genotypic relationships in *L. monocytogenes* isolated from fresh produce (Smith *et al.* 2019). The presence of virulence genes was found in all isolates, and no gene was specific to a single category of origin.

The work presented in this thesis probed beyond phenotypic screening, comparing the whole genomes of these isolates to those of meat and clinical origin using genome wide association. This assessed the genetic relatedness of these isolates and employed statistical methods to mine for genes belonging to a specific category of origin. It was found that there were 218 CDS associated with vegetable origin, 11 CDS associated with clinical origin and zero associated with meat. Many of these coded for proteins of unknown function, as found by similar studies previously (Pirone-Davies *et al.* 2018), highlighting the need for further research in this area to elucidate the function of these proteins. The work undertaken in chapter 4 is the first to examine the genome of *L. monocytogenes* using GWAS to extract potential genes that effect the survival of *L. monocytogenes* within the fresh produce microbiome.

(iv) Assess the incidence of AMR-associated genes in foodborne microbes.

The presence of genes associated with AMR was screened in the microbiome associated with fresh produce in chapter 3 of this thesis. This is the first known study utilising NGS to assess the burden of AMR-associated genes in fresh produce from the UK food supply chain. This study found evidence of the presence of four AMR-associated genes: *H-NS, CRP, mexF* and *mexB*. These resistance genes are potentially of clinical importance, encoding for gene regulators and efflux pumps, but due to limitations meaning metatranscriptomics cannot assign the genes to the member of the microbiome the full clinical impact cannot be ascertained. Previous studies have examined bacteria isolated from food (Holzel *et al.* 2018) or the general plant microbiome (Fogler *et al.* 2019) for AMR or AMR genes, but these do not reflect the burden of AMR associated with food as it would reach the consumer, and therefore the potential for these genes to be passed to bacteria within the gut microbiome. The current study, by focusing on ready to eat products, demonstrates the burden of AMR potentially entering the gut on foods. The gut has previously been shown to denote one of the best ecological niches for horizontal gene transfer (Lerner *et al.* 2017), and it has been noted that these genes could pass from the food microbiome to the gut microbiome (Rolain 2013; Aarts and Margolles 2015).

Isolates of *L. monocytogenes* isolated from the fresh produce supply chain were additionally screened for AMR associated genes in chapter 4 of this thesis. This built on previous work screening *L. monocytogenes* for antimicrobial resistance genes (Painset *et al.* 2019; Smith *et al.* 2019). It was found, in concordance with previous work, that there were low levels of AMR genes found in *L. monocytogenes* isolated from food and clinical cases the UK.

Can next generation sequencing be used as a screening tool for human pathogenic organisms in fresh produce?

As identified in chapter 2, the limit of detection (LoD) of current NGS methodologies on the MiSeq are not sensitive enough to detect the legislative limits of human pathogens associated with fresh produce (Health Protection Agency (now Public Health England) 2009). These limits vary from pathogen to pathogen, ranging from not detected in 25 g for pathogens including *Salmonella* to 10^2 CFU in 25 g for *L. monocytogenes*. It therefore is not suitable as a screening tool to detect human pathogens within the fresh produce microbiome as the LoD was found to be 10^4 CFU of *Salmonella* and 10^5 PFU of MS2 per sample. The utilisation of the high throughput Illumina platforms, such as the NovaSeq or HiSeq, would allow for a lower cost per base, but are still too expensive for routine use and additionally would yield shorter read lengths. New technologies are available, for example the PromethION by Oxford Nanopore, which have a lower cost per base than the Illumina platforms, allowing a greater depth of sequencing and so increasing sensitivity. While the error rate of these technologies is currently a lot higher than that for the Illumina platforms (Lima *et al.* 2019), the longer read length may allow for a more accurate classification of the microbiome despite this error rate (Pearman *et al.* 2019). The work undertaken in chapter 2 needs to be extended to the assessment and development of these new platforms as a first step in any validation procedure to allow the identification of the LoD and examine the effect the higher error rate of the Oxford Nanopore platforms has on the specificity and LoD.

Future work will need to focus on the delivery of a more representative nucleic acid sample to the sequencing instrument. This should facilitate increased sensitivity, and decreased cost. The extraction methodology used in chapters 2 and 3 was designed to extract the total microbiome, including internalised and strongly adhered members of the community, and to mimic the stomaching method utilised as part of the microbiological testing undertaken. This

method, while allowing for examination of the total microbiome and thereby not underestimating the LoD, may have increased the proportion of the extracted sample belonging to the plant matrix (in the case of chapter 2 this was lettuce). Further work therefore must be undertaken to create an extraction method that allows for extraction of the total microbiome while minimising contamination from the plant matrix. Some key methods to examine may include, centrifugation of the samples to remove the plant matrix and concentrate the microbiome or washing the microbiome from the fresh produce and extracting this instead of a homogenate (Ahn *et al.* 2012; Fouhy *et al.* 2016). Post extraction molecular enrichment methods were examined in chapter 2 and, while the best method was ribosomal depletion, data still yielded a high percentage RNA from the plant matrix. New methods enabling molecular enrichment may allow for greater sequencing depth to represent the microbiome. One method worthy of exploration may be the utilisation of nucleic acid probes (Shih *et al.* 2018). These can be used to target either the microbiome or the plant matrix, delivering a higher proportion of nucleic acid extract pertaining to the microbiome, in a similar way to ribosomal depletion. The probes would need to be less specific than those for ribosomal depletion. This may not be a viable solution for screening methods, as it is costly to design new probes, which would be needed for each plant matrix, and there would remain a trade-off between specificity, and the proportion of plant matrix removed. An additional new technology that could be used as a method for molecular probing is the CRISPR/Cas9 system. CAS-9 probes could be designed to specifically target plant matrix sequences and break them, thereby decreasing the likelihood of them making it to sequencing. Initial work using this system has shown promise, but has primarily focused on 16S where the contamination that requires removal is limited to ribosomal and chloroplast RNA (Song and Xie 2020). Significant further research may be needed to apply this to metagenomics/metatranscriptomics studies in the future.

A further issue with the use of NGS as a screening tool for human pathogenic organisms is that there is often bias or misassignment attributed to the bioinformatics protocol and databases used. This was seen in chapter 2, where several methods of data analysis showed *Salmonella* in the results in samples where no *Salmonella* was detected using qPCR. When these reads were mapped to the Salmonella genome (data not shown) it was found that the reads were predominantly in the rRNA region, and when mapped to the lettuce chloroplast sequence they mapped well with the chloroplast RNA. This was notably true for 16S

sequencing where high proportions of the reads were filtered out in chapter 2 by mapping to the chloroplast. Without this mapping, higher levels of misassignment was also seen. This is a notable issue in the sequencing of fresh produce, where most extraction methods will extract chloroplast RNA as easily as bacterial RNA. This has an impact on the use of NGS as a screening tool for fresh produce as even low levels of misassignment when using NGS as a screening tool could give a false positive result, a potentially serious and costly problem that would require further testing to ascertain whether it was a true positive. Additionally, the bias in database and bioinformatics tools used would lead to the results being incomparable between laboratories, therefore before these methods could be used as a true screening tool standardised methods, as with the ISO methods currently used, should be validated and verified, and adopted by all laboratories. This would allow for quantification of the uncertainties and issues with the techniques, allowing the users to know information on the detection limits, potential for false negatives and false positives and the standardisation and comparability of results between laboratories.

Another reason that NGS is not yet applicable as a screening method for the presence of human pathogens in fresh produce is that, as with all molecular methods, these techniques have difficulties in determining whether the microorganisms found are viable or infective (Emerson *et al.* 2017). This is reflected in chapter 3 where, when NGS is compared to microbiological methods the data for NGS show a high false positive rate for *Listeria* species. Additionally, RNA based methods, as utilised in chapter 2 and 3, may give a more accurate representation of the viability of the sequenced organisms than DNA based methods, as RNA will potentially only be available from actively transcribing members of the microbiome and, being shorter lived than DNA, could allow for a better correlation with viability. However, there has been little or no research into whether presence of RNA correlates with viability, and in chapter 3 of this thesis data from RNA methods did not correlate with microbiological results. The divergence found in chapter 3 could be due to the presence of viable but non-culturable strains potentially missed by microbiological screens (Highmore *et al.* 2018), although the role this plays in food microbiology is highly debated by many in the field. In conclusion, further work is needed to validate RNA sequencing as a potential methodology and prove the links to the viability of the organisms it detects.

Are NGS methods suitable to examine phenotypic traits, such as AMR?

NGS is often used for the screening of whole genome and metagenomics or metatranscriptomics data for genes associated with phenotypes of interest, such as antimicrobial resistance (AMR), as undertaken in chapters 3 and 4, or virulence, as undertaken in chapter 4. This is a key benefit of the use of NGS as it allows us to gain more insight into the genes associated with fresh produce samples.

One issue with the screening of NGS data for genes associated with specific phenotypic traits is that currently there is limited knowledge as to whether the presence and absence of genes associated with these phenotypes correlate with the phenotype itself. In the context of AMR, many studies have utilised NGS to screen for AMR associated genes (Cerqueira *et al.* 2019) but few studies have yet to link these data to conventional phenotypic screen data. Those studies that have been undertaken found that, for isolates that have undergone phenotypic screening, WGS screening for AMR associated genes finds AMR genes in most, but not all isolates of a resistant phenotype (Neuert *et al.* 2018; Guo *et al.* 2019). The lack of concordance between NGS and phenotypic screen may be due to the limited databases of AMR associated genes, and with further exploration and the detection of novel AMR genes this correlation should get stronger. In addition, the screening of AMR genes in microbiome samples is problematic as the data does not give insights into which organism the AMR genes found were associated with. As AMR genes may be silent in some hosts this is a key problem with the current methods (Dantas and Sommer 2012). Long read technologies, such as the Oxford Nanopore sequencers, may allow us to put these genes in genomic context (Ashton *et al.* 2014), although as many AMR genes are plasmid borne current techniques will not be able to solve the issues for all genes.

The presence of a gene also does not mean phenotypic activity within the ecosystem, a problem associated with phenotypic screening methods as well. The use of RNA based methods, such as in chapter 3, may provide more biologically relevant data as it will only detect genes that are actively transcribed (Bashiardes *et al.* 2016). For the case of AMR associated genes, active transcription may point to the presence of antimicrobial compounds, therefore these organisms may be under greater selection pressure, increasing the likelihood of transmission of AMR genes to other organisms. Using RNA based methods to screen *L. monocytogenes* isolates (for example those analysed using WGS in chapter 4) under different conditions associated with fresh produce, for example at cool temperatures

and on plant-based media, could yield data on differential gene expression in the fresh produce production chain versus under standard laboratory conditions. This may lead to greater information on genes associated with survival in these conditions and unlike WGS data, where the information is on presence and absence of genes, RNAseq will allow the analysis of actively transcribing genes.

Is NGS usable as a tool for the assessment of the microbiological safety of fresh produce?

In its current form NGS cannot fully replace current methodologies. Microbiome analysis is not sensitive enough for the detection of human pathogens in fresh produce (as outlined in chapter 2), but in food spoilage studies where the concentration of organisms of interest is higher, this LoD is sufficient to identify organisms that may lead to spoilage, as found in chapter 3. In chapter 3, NGS was successfully used in conjunction with other metadata, including microbiological results regarding the presence and absence of human pathogens or indicators, to allow for the screening of the microbiome. Several members of the microbiome were found to positively correlate with the presence of Enterobacteriaceae and *Listeria* species. With further study, examining additional samples including those positive for human pathogens, these could be utilised as indicator organisms, or potentially included as a risk factor associated with the survival of human pathogens within the fresh produce microbiome, making the study of these even more important.

Pseudonocardia was inversely correlated with presence of Enterobacteriaceae, and future work could be undertaken to examine the potential use of this organism as a biocontrol agent to decrease the survival of human pathogens on fresh produce. Initial work will need to be done to ascertain that the organism would be safe for human consumption although it has been previously reported as a member of both the fresh produce microbiome (Cerqueira *et al.* 2019), including in chapter 3, and the human microbiome (Bassiouni *et al.* 2015). Further work will need to be undertaken, likely in the form of a glass house trial, to show that *Pseudonocardia* can survive, and potentially reproduce, in the fresh produce microbiome. It would also be important to demonstrate this organism is able to confer a negative effect on the survival of human pathogens. This could be done through the direct application of the bacterium, or via screening of these bacteria for production of antibiotic compounds that can then be screened for activity against human pathogenic organisms, the latter also allowing for the potential discovery of novel antimicrobials of therapeutic interest.

One of the key issues with use of NGS to examine the microbiome, notably using metatranscriptomics, is the lack of robust analysis pipelines and well curated sequence databases, which in turn lead to the misassignment of sequences and misrepresentation of the microbiome. This was shown in chapter 3, where the prevalence of *Alteromonas* in the samples is likely in part to be due to misassignment. Time needs to be invested in the creation of databases with well curated, accurate and appropriate representation of species for NGS analysis of the fresh produce microbiome. This microbiome is diverse and, compared with the human microbiome, poorly studied and therefore the current databases are biased against this environment which likely leads to misassignment. Bioinformatic pipelines also need to be created and validated on mock communities or simulated fresh produce microbiomes to allow for confidence in the microbiome assignment of bioinformatic tools.

The microbiome data produced in chapter 3 was successfully screened for AMR associated genes. This is important as it may allow for the qualification of the burden of AMR within the food supply chain, and further interrogation of the data may also allow for the identification of genes encoding for novel antimicrobial compounds. AMR screening of *L. monocytogenes* isolates was also successfully undertaken in chapter 4, giving us further information on the risk of AMR in a medically important bacterium in the UK. Further work could be done on these data to screen for novel AMR genes (Adu-Oppong *et al.* 2016) in which the utilisation of machine learning and newly emerging computer based methodologies may be extremely useful and could allow for the discovery of many new genes of interest (Arango-Argoty *et al.* 2018).

Chapter 4 also utilised NGS, through WGS of *L. monocytogenes*, to screen *L. monocytogenes* for specific genes of interest, including AMR, virulence and, in conjunction with metadata on the origin of the isolate, genes associated with survival in the fresh produce environment. Many genes of potential interest were correlated with isolates from fresh produce, but more work is needed to establish what effect, if any, these may be having on survival of *L. monocytogenes*. The first step in this could be to increase the power of statistical inferences by increasing the number of isolates screened. This would likely lead to a smaller number of genes being identified as important and allow for more formal screening of their potential functions. The functional screens would allow for the identification of genes essential for

survival, which may be targeted to prevent the persistence of these isolates in the fresh produce environment.

The use of WGS of pathogens isolated using traditional microbiological methods is now implemented as part of outbreak tracking within the UK (Public Health England 2019) but is currently too expensive to be routinely deployed in non-outbreak or industrial settings. New technologies, such as the Oxford Nanopore sequencers, may decrease this cost, but at present cannot obtain the resolution and accuracy needed to source track outbreaks. As further advances are made in this technology, decreasing the error rates, this may become a cheaper alternative to existing methods allowing for greater use of WGS. There is also a requirement in this field for more standardised pipelines for the analysis of WGS data, and validation of these pipelines, to allow for the comparison of results across labs and to ensure the results obtained are true. This is notably important in outbreak tracking, where there is a potential burden of fault with a commercial company who are the source of a product that has led to an outbreak.

In addition, data generated by NGS currently requires intensive computer infrastructure and bioinformatic knowledge to analyse. The generation of standardised and potentially centralised facilities, that can be used by laypeople to analyse data, is important if NGS is to be used more routinely in the food microbiology setting.

What are the technological advances required for the routine use of NGS in the assessment of microbiological safety of fresh produce?

As outlined, there are many technological advances that are required to overcome the current limitations in the use of NGS within food microbiology. These include but are not limited to:

1. Improvement of methods of extraction and molecular enrichment to allow sequencing of the total microbiome while minimising plant matrix contamination.

The method of extraction and enrichment plays a large role in the microbiome sequenced and the proportion of plant matrix contamination sequenced (Fricker *et al.* 2019). Many enrichment methods have focused on depletion of human or mouse associated nucleic acids due to the focus of research in these areas (Heravi *et al.* 2020). Existing technologies for depletion of plant matrix contamination have a limited success, as seen in chapter 2 of this thesis, therefore there is a focus on potential novel technologies to be utilised in this area.

Techniques such as CRISPR-Cas (Song and Xie 2020), CRISPR-Cap (Lee *et al.* 2019) and aptamer technologies (Sun and Zu 2015), all promise to deliver a more targeted depletion, but as yet are in the early stages of their development.

2. The assessment of methods to judge their ability to determine viability of targets, notably for human pathogen screening

The use of molecular methods to determine viability of the targets has been a challenge to molecular methodologies. It has been previously examined, predominantly in the context of PCR based methods (Reyneke *et al.* 2017) and in the context of food microbiology (Dinh Thanh *et al.* 2017; Agusti *et al.* 2018). Despite this, there has been limited success in the use of these methods, with studies showing potential false positives or biases introduced by current methodologies (Agusti *et al.* 2017; Codony *et al.* 2020). Additionally, none of these methods have been examined in the context of NGS. It is therefore important that these methods are assessed in the context of human pathogens in the fresh produce microbiome when assessed using NGS.

3. Lower cost sequencing per base to a greater depth of sequencing, leading to an LoD in the range of other methods, attainable at a cost low enough to make this screening routine

The cost of NGS has decreased dramatically since the introduction of these technologies. It is now in some applications commercially viable to employ NGS methods over traditional microbiological screening (Torchia *et al.* 2019), but this is dependent on the cost, sensitivity and specificity of the microbiological methods currently used and in most cases NGS is still deemed more costly than traditional techniques (Gwinn *et al.* 2019). In the context of food microbiology although there has been much research into the potential applications of NGS in this field, the cost often prohibits utilisation of NGS in a routine fashion (Jagadeesan *et al.* 2019). With new technological advances, such as the Promethion by Oxford Nanopore, we are again seeing a decrease in the cost of NGS. But further technologies and validation of procedures on these platforms are needed to determine their viability as screening tools.

4. Standardised analysis pipelines that do not require local computing infrastructures and expertise

As shown in chapter 2 of this thesis, the choice of bioinformatics technique affects the results of NGS sequencing, highlighting the need for standardised analysis techniques that have been validated. In addition to this, current bioinformatics processes often require access to complex computer clusters and technical knowledge to run and manipulate, making the analysis costly and often unique to each laboratory. Many consortiums have been established to try and create consensus in the analysis of NGS data between institutions (e.g. International Cancer Genome Consortium, the External RNA Control Consortium, the Genome in a Bottle Consortium, Earth Microbiome Project), but these are primarily focused on clinical applications.

5. Curated and appropriate sequence databases

The sequencing database, in addition to the laboratory and analysis protocols, can also influence the data obtained from NGS, including the rates of misassignment and the perceived diversity of samples (Allali *et al.* 2017). As seen in chapter 2, the lack of relevant data in the database (in this case the lack of the full lettuce genome) can lead to misassignment of reads to incorrect taxa. Errors in databases can additionally introduce errors into results. It is therefore key to have a centrally agreed and well curated database with a broad scope, including potential matrices, to allow for the qualification of the bias associated with these databases and to decrease the potential for misassignment.

6. Fully validated processes to allow for inter-laboratory consistency of results

The key to use of any technology as a recognised screening tool is the presence of fully validated processes, with data available on the detection limits, scope of the method, sensitivity and specificity of the technique. The ISO standard protocols are a requirement for most commonly tested for food-borne pathogens, and therefore this level of between laboratory comparability would be needed for NGS to allow this to be fully adopted as a routine screening tool in the context of food microbiology.

The above technological advances, and appropriate validation procedures would allow for a routine use of NGS in the assessment of microbiological safety of fresh produce. The focus

needs to be on the validation of any new technologies to allow for the presentation of accurate and repeatable results, fundamental for risk management in the food supply chain.

What future impact could NGS have in the assessment of microbiological safety of fresh produce?

NGS is likely to be an important tool in the future for the assessment of the microbiological safety of fresh produce. It may be a generalised method for the screening of all potential human pathogenic organisms, their associated virulence (and therefore whether they would be a risk to consumers), and the presence of AMR (and therefore the levels of AMR that consumers are exposed to). This would negate the need for multiple tests and the high cost, or high risk associated with under or over screening of fresh produce for human pathogens using traditional targeted methodologies. Although this is currently not feasible, NGS can be used in conjunction with other data, as in chapter 3, to allow for the discovery of new indicator species, or a panel of species, that better correlate with presence of human pathogens. This would allow for wider testing, and decreased risk of false negatives or false positives, of fresh produce contamination in the food chain, so decreasing both grower and consumer risk.

The large quantity of information within microbiome studies may allow for the discovery of novel biocontrol agents, such as *Pseudonocardia* described in chapter 3, that can be applied in the field or processing environment to prevent the survival or, notably in the case of *L. monocytogenes*, growth of human pathogens on fresh produce. These data, as well as WGS data, may also allow for the discovery of novel biocides or antibiotics produced within these organisms. WGS may lead to the identification of novel pathways required for the survival of human pathogens within the fresh produce microbiome. These could then be targeted in the design of novel biocides.

The advances in NGS technologies are rapid, with new platforms, sequencing kits and bioinformatic analysis techniques being designed at a pace quicker than any formal validation procedure can currently keep up with. It is important that validation and verification of these are undertaken before use, and that validation bodies are pragmatic in the accreditation of these techniques, as a non-targeted approach does not necessarily fit into the current scheme of validation. Although this study has shown that NGS cannot currently be routinely used in food microbiology, there is no reason that in the future, given

the technological advances outlined, NGS will not be a standard part of the assessment of microbiological risk within the fresh produce supply chain.

References

Aarts H, Margolles A (2015) Antibiotic resistance genes in food and gut (non-pathogenic) bacteria. Bad genes in good bugs. Frontiers in Microbiology 5:Article 754.

Adams IP, Glover RH, Monger WA, Mumford R, Jackeviciene E, Navalinskiene M, Samuitiene M, Boonham N (2009) Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. Molecular Plant Pathology 10:537-545.

Adams MH (1959) Bacteriophages. Interscience Publishers Ltd., London.

Adu-Oppong B, Gasparrini AJ, Dantas G (2016) Genomic and functional techniques to mine the microbiome for novel antimicrobials and antimicrobial resistance genes. Annals of the New York Academy of Sciences:1-17.

Agilent Technologies (2015) Agilent High Sensitivity D1000 ScreenTape System Quick Guide G2964-90131 rev. D. 09/2015 edn.

Aguado V, Vitas AI, Garcia-Jalon I (2004) Characterization of *Listeria monocytogenes* and *Listeria innocua* from a vegetable processing plant by RAPD and REA. International Journal of Food Microbiology 90:341-347.

Agusti G, Fittipaldi M, Codony F (2017) False-Positive Viability PCR Results: An Association with Microtubes. Current Microbiology 74:377-380.

Agusti G, Fittipaldi M, Codony F (2018) Optimization of a Viability PCR Method for the Detection of Listeria monocytogenes in Food Samples. Current Microbiology 75:779-785.

Ahn J-H, Kim B-Y, Song J, Weon H-Y (2012) Effects of PCR cycle number and DNA polymerase type on the 16S rRNA gene pyrosequencing analysis of bacterial communities. Journal of Microbiology 50:1071-1074.

Allali I, Arnold JW, Roach J, Cadenas MB, Butz N, Hassan HM, Koci M, Ballou A, Mendoza M, Ali R et al. (2017) A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. BMC Microbiology 17:194-210.

Allard MW, Luo Y, Strain E, Cong L, Keys C, Son I, Stones R, Musser SM, Brown EW (2012) High resolution clustering of *Salmonella enterica* serovar Montevideo strains using a next-generation sequencing approach. BMC Genomics 13:32-50.

Amrutha B, Sundar K, Shetty PH (2017) Study on *E. coli* and *Salmonella* biofilms from fresh fruits and vegetables. Journal of Food Science and Technology 54:1091-1097.

Andreevskaya M, Jääskeläinen E, Johansson P, Ylinen A, Paulin L, Björkroth J, Auvinen P (2018) Food spoilage-associated *Leuconostoc*, *Lactococcus*, and *Lactobacillus* species display different survival strategies in response to competition. Applied Environmental Microbiology 84:e00554-00518.

Arango-Argoty G, Garner E, Pruden A, Heath LS, Vikesland P, Zhang L (2018) DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome 6.

Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, Wain J, O'Grady J (2014) MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nature Biotechnology 33:296-300.

Asselin JE, Eikemo H, Perminow J, Nordskog B, Brurberg MB, Beer SV (2019) *Rahnella* spp. are commonly isolated from onion (*Allium cepa*) bulbs and are weakly pathogenic. Journal of Applied Microbiology 127:812-824.

Aureli P, Ferrini AM, Mannoni V, Hodzic S, Wedell-Weergaard C, Oliva B (2003) Susceptibility of *Listeria monocytogenes* isolated from food in Italy to antibiotics. International Journal of Food Microbiology 83:325-330.

Aw TG, Wengert S, Rose JB (2016) Metagenomic analysis of viruses associated with field-grown and retail lettuce identifies human and animal viruses. International Journal of Food Microbiology 223:50-56.

Ayrapetyan M, Oliver JD (2016) The viable but non-culturable state and its relevance in food safety. Current Opinion in Food Science 8:127-133.

Bah A, Albano H, Barbosa JB, Fhoula I, Gharbi Y, Najjari A, Boudabous A, Teixeira P, Ouzari HI (2019) Inhibitory Effect of *Lactobacillus plantarum* FL75 and *Leuconostoc mesenteroides* FL14

against Foodborne Pathogens in Artificially Contaminated Fermented Tomato Juices. BioMed Research International 2019:6937837.
Balcazar JL, Subirats J, Borrego CM (2015) The role of biofilms as environmental reservoirs of antibiotic resistance. Frontiers in Microbiology 6:1216.
Barak JD, Kramer LC, Hao Ly (2010) Colonization of tomato plants by *Salmonella enterica* is cultivar dependent, and type 1 trichomes are preferred colonization sites. Applied and Environmental Microbiology 77:498-504.
Bartsch C, Höper D, Mäde D, Johne R (2018) Analysis of frozen strawberries involved in a large norovirus gastroenteritis outbreak using next generation sequencing and digital PCR. Food Microbiology 76:390-395.
Bartz JA, Yuk H-G, Mahovic MJ, Warren BR, Sreedharan A, Schneider KR (2015) Internalization of *Salmonella enterica* by tomato fruit. Food Control 55:141-150.
Bashiardes S, Zilberman-Schapira G, Elinav E (2016) Use of metatranscriptomics in microbiome research. Bioinformatics and Biology Insights 10:19-25.
Bassiouni A, Cleland EJ, Psaltis AJ, Vreugde S, Wormald P-J (2015) Sinonasal Microbiome Sampling: A Comparison of Techniques. PloS One 10.
Beckers B, Op De Beeck M, Thijs S, Truyens S, Weyens N, Boerjan W, Vangronsveld J (2016) Performance of 16s rDNA primer pairs in the study of rhizosphere and endosphere bacterial microbiomes in metabarcoding studies. Frontiers in Microbiology:https://doi.org/10.3389/fmicb.2016.00650.
Bell RL, Jarvis KG, Ottesen AR, McFarland MA, Brown EW (2016) Recent and emerging innovations in *Salmonella* detection: a food and environmental perspective. Microbial Biotechnology 9:279-292.
Benítez-Páez A, Portune KJ, Sanz Y (2016) Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. GigaScience 5:4.
Bergholz TM, Moreno Switt AI, Wiedmann M (2014) Omics approaches in food safety: fulfilling the promise? Trends in Microbiology 22:275-281.
Beuchat LR, Ryu J-H (1997) Produce handling and processing practices. Emerging Infectious Diseases 3:459-465.
Black JS, Salto-Tellez M, Mills KI, Catherwood MA (2015) The impact of next generation sequencing technologies on haematological research – A review. Pathogenesis 2:9-16.
Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F et al. (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology 37:852-857.
Bonasera JM, Asselin JAE, Beer SV (2017) Lactic Acid Bacteria Cause a Leaf Blight and Bulb Decay of Onion (*Allium cepa*). Plant Disease 101:29-33.
Bonsaglia ECR, Silva NCC, Fernades Júnior A, Araújo Júnior JP, Tsunemi MH, Rall VLM (2014) Production of biofilm by *Listeria monocytogenes* in different materials and temperatures. Food Control 35:386-391.
Borucki MK, Peppin JD, White D, Loge F, Call DR (2003) Variation in biofilm formation among strains of *Listeria monocytogenes*. Applied and Environmental Microbiology 69:7336-7342.
Boyle EC, Bishop JL, Grassl GA, Finlay BB (2007) *Salmonella*: from pathogenesis to therapeutics. Journal of Bacteriology 189:1489-1495.
Brandl MT (2008) Plant lesions promote the rapid multiplication of *Escherichia coli* O157:H7 on postharvest lettuce. Applied and Environmental Microbiology 74:5285-5289.
Brandl MT, Amundson R (2008) Leaf age as a risk factor in contamination of lettuce with *Escherichia coli* O157:H7 and *Salmonella enterica*. Applied and Environmental Microbiology 74:2298-2306.
Bridier A, Sanchez-Vizuete P, Guilbaud M, Piard JC, Naitali M, Briandet R (2015) Biofilm-associated persistence of food-borne pathogens. Food Microbiology 45:167-178.
Buchanan RL, S.G. E, Miller RL, Sapers GM (1999) Contamination of intact apples after immersion in an aqueous environment containing *Escherichia coli* O157:H7. Journal of Food Protection 62:444-450.

Buchholz U, Bernard H, Werber D, Böhmer MM, Remschmidt C, Wilking H, Deleré Y, an der Heiden M, Adlhoch C, Dreesman J et al. (2011) German outbreak of *Escherichia coli* O104:H4 associated with sprouts. New England Journal of Medicine 365:1763-1770.

Burnett SL, Chen J, Beuchat LR (2000) Attachment of *Escherichia coli* O157:H7 to the surfaces and internal structures of apples as detected by confocal laser microscopy. Applied and Environmental Microbiology 66:4679-4687.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Cao Y, Fanning S, Proos S, Jordan K, Srikumar S (2017) A Review on the applications of Next Generation Sequencing technologies as applied to food-related microbiome studies. Frontiers in Microbiology 8:http://doi.org/10.3389/fmicb.2017.01829.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI et al. (2010) QIIME allows analysis of highthroughput community sequencing data. Nature Methods 7:335-336.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M et al. (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME Journal 6:1621-1624.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proceedings of the National Academy of Sciences 108:4516-4522.

Carpentier B, Chassaing D (2004) Interactions in biofilms between *Listeria monocytogenes* and resident microorganisms from food industry premises. International Journal of Food Microbiology 97:111-122.

Cernava T, Erlacher A, Soh J, Sensen CW, Grube M, Berg G (2019) Enterobacteriaceae dominate the core microbiome and contribute to the resistome of arugula (*Eruca sativa* Mill.). Microbiome 7:13.

Cerqueira F, Matamoros V, Bayona J, Pina B (2019) Antibiotic resistance genes distribution in microbiomes from the soil-plant-fruit continuum in commercial *Lycopersicon esculentum* fields under different agricultural practices. Science of the Total Environment 652:660-670.

Chan YC, Wiedmann M (2009) Physiology and genetics of *Listeria monocytogenes* survival and growth at cold temperatures. Critical Reviews in Food Science and Nutrition 49:237-253.

Chang Y-F, Marvasi M, Hochmuth GJ, Giurcanu MC, George AS, Noel JT, Bartz J, Teplitski M (2013) Factors that affect proliferation of *Salmonella* in tomatoes post-harvest: the roles of seasonal effects, irrigation regime, crop and pathogen genotype. PloS One 8:e80871.

Chapman SJ (1990) *Thiobacillus* populations in some agricultural soils. Soil Biology and Biochemistry 22:479-482.

Chen Y, Gonzalez-Escalona N, Hammack TS, Allard MW, Strain EA, Brown EW (2016) Core genome multilocus sequence typing for identification of globally distributed clonal groups and differentiation of outbreak strains of *Listeria monocytogenes*. Applied and Environmental Microbiology 82:6258-6272.

Chibani-Chennoufi S, Bruttin A, Dillmann ML, Brussow H (2004) Phage-host interaction: an ecological perspective. Journal of Bacteriology 186:3677-3686.

Cleveland J, Montville TJ, Nes IF, Chikindas ML (2001) Bacteriocins: safe, natural antimicrobials for food preservation. International Journal of Food Microbiology 71:1-20.

Codex Alimentarius Commission (Adopted 2003. Revision 2010 (new Annex III for Fresh Leafy Vegetables), 2012 (new Annex IV for Melons), 2013 (new Annex V for Berries)) Code of hygienic practice for fresh fruits and vegetables.

Codony F, Dinh-Thanh M, Agusti G (2020) Key Factors for Removing Bias in Viability PCR-Based Methods: A Review. Current Microbiology 77:682-687.

Condell O, Iversen C, Cooney S, Power KA, Walsh C, Burgess C, Fanning S (2012) Efficacy of biocides used in the modern food industry to control *Salmonella enterica*, and links between biocide tolerance and resistance to clinically relevant antimicrobial compounds. Applied and Environmental Microbiology 78:3087-3097.

Conter M, Paludi D, Zanardi E, Ghidini S, Vergara A, Ianieri A (2009) Characterization of antimicrobial resistance of foodborne *Listeria monocytogenes*. International Journal of Food Microbiology 128:497-500.
Cook N, Knight A, Richards GP (2016) Persistence and elimination of human norovirus in food and on food contact surfaces: a critical review. Journal of Food Protection 79:1273-1294.
Cook N, Williams L, D'Agostino M (2019) Prevalence of Norovirus in produce sold at retail in the United Kingdom. Food Microbiology 79:85-89.
Creel RH (1912) Vegetables as a possible factor in the dissemination of Typhoid fever. Public Health Reports 27:187-193.
Critzer FJ, Doyle MP (2010) Microbial ecology of foodborne pathogens associated with produce. Current Opinion in Biotechnology 21:125-130.
Daba H, Pandian S, Gosselin JF, Simard RE, Huang J, Lacroix C (1991) Detection and activity of a bacteriocin produced by Leuconostoc mesenteroides. Applied and Environmental Microbiology 57:3450-3455.
Dailey RC, Welch LJ, Hitchins AD, Smiley RD (2015) Effect of *Listeria seeligeri* or *Listeria welshimeri* on *Listeria monocytogenes* detection in and recovery from buffered *Listeria* enrichment broth. Food Microbiology 46:528-534.
Dantas G, Sommer MO (2012) Context matters - the complex interplay between resistome genotypes and resistance phenotypes. Current Opinion in Microbiology 15:577-582.
de Boer P, Caspers M, Sanders J-W, Kemperman R, Wijman J, Lommerse G, Roeselers G, Montijn R, Abee T, Kort R (2015) Amplicon sequencing for the quantification of spoilage microbiota in complex foods including bacterial spores. Microbiome 3:1-30.
de Vasconcelos Byrne V, Hofer E, Vallim DC, de Castro Almeida RC (2016) Occurrence and antimicrobial resistance patterns of *Listeria monocytogenes* isolated from vegetables. Brazilian Journal of Microbiology 47:438-443.
Deering AJ, Mauer LJ, Pruitt RE (2012) Internalization of *E. coli* O157:H7 and *Salmonella* spp. in plants: A review. Food Research International 45:567-575.
Delcour AH (2009) Outer membrane permeability and antibiotic resistance. Biochimica et Biophysica Acta 1794:808-816.
den Bakker HC, Cummings CA, Ferreira V, Vatta P, Orsi RH, Degoricija L, Barker M, Petrauskene O, Furtado MR, Wiedmann M (2010) Comparative genomics of the bacterial genus *Listeria*: Genome evolution is characterized by limited gene acquisition and limited gene loss. BMC Genomics 11:668-688.
Department for Environment Food and Rural Affairs (2015) The guide to cross compliance in England.
DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL (2006) Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Applied and Environmental Microbiology 72:5069-5072.
Dinh Thanh M, Agusti G, Mader A, Appel B, Codony F (2017) Improved sample treatment protocol for accurate detection of live Salmonella spp. in food samples by viability PCR. PloS One 12:e0189302.
Djordjevic D, Wiedmann M, McLandsborough LA (2002) Microtiter plate assay for assessment of *Listeria monocytogenes* biofilm formation. Applied and Environmental Microbiology 68:2950-2958.
Domesle KJ, Yang Q, Hammack TS, Ge B (2018) Validation of a *Salmonella* loop-mediated isothermal amplification assay in animal food. International Journal of Food Microbiology 264:63-76.
Ellington MJ, Ekelund O, Aarestrup FM, Canton R, Doumith M, Giske C, Grundman H, Hasman H, Holden MTG, Hopkins KL et al. (2017) The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. Clinical Microbiology and Infection 23:2-22.
Emerson JB, Adams RI, Roman CMB, Brooks B, Coil DA, Dahlhausen K, Ganz HH, Hartmann EM, Hsu T, Justice NB et al. (2017) Schrodinger's microbes: Tools for distinguishing the living from the dead in microbial ecosystems. Microbiome 5:86.

Endrullat C, Glökler J, Franke P, Frohme M (2016) Standardization and quality management in next-generation sequencing. Applied & Translational Genomics 10:2-9.
Epicentre (2013) ScriptSeq™ Complete Kit (Plant Seed/Root) Lit. # 349 4/2013 EPILIT349 Rev. A.
Ercolini D, Ferrocino I, Nasi A, Ndagijimana M, Vernocchi P, La Storia A, Laghi L, Mauriello G, Guerzoni ME, Villani F (2011) Monitoring of microbial metabolites and bacterial diversity in beef stored under different packaging conditions. Applied and Environmental Microbiology 77:7372-7381.
Escolar C, Gomez D, Del Carmen Rota Garcia M, Conchello P, Herrera A (2017) Antimicrobial resistance profiles of *Listeria monocytogenes* and *Listeria innocua* isolated from ready-to-eat products of animal origin in Spain. Foodborne Pathogens and Disease 14:357-363.
European Food Safety Authority (2016) The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2015. EFSA Journal 14.
European Food Safety Authority (2018) The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2017. EFSA Journal 16.
Fatica MK, Schneider KR (2011) *Salmonella* and produce: Survival in the plant environment and implications in food safety. Virulence 2:573-579.
Ferreira V, Wiedmann M, Teixeira P, Stasiewicz MJ (2014) *Listeria monocytogenes* persistence in food-associated environments: epidemiology, strain characteristics, and implications for public health. Journal of Food Protection 77:150-170.
Fogler K, Guron GKP, Wind LL, Keenum IM, Hession WC, Krometis L-A, Strawn LK, Pruden A, Ponder MA (2019) Microbiota and antibiotic resistome of lettuce leaves and radishes grown in soils receiving manure-based amendments derived from antibiotic-treated cows. Frontiers in Sustainable Food Systems 3.
Food Standards Agency (2011) Foodborne Disease Strategy 2010-15.
Forsberg KJ, Reyes A, Wang B, Selleck EM, Sommer MOA, Dantas G (2012) The shared antibiotic resistome of soil bacteria and human pathogens. Science 337:1107-1111.
Fouhy F, Clooney AG, Stanton C, Claesson MJ, Cotter PD (2016) 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. BMC Microbiology 16:123-136.
Fox EM, Allnutt T, Bradbury MI, Fanning S, Chandry PS (2016) Comparative genomics of the *Listeria monocytogenes* ST204 subgroup. Frontiers in Microbiology 7:2057.
Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M, Carric JA (2012) PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. BMC Bioinformatics 13:87.
Franz E, Delaquis P, Morabito S, Beutin L, Gobius K, Rasko DA, Bono J, French N, Osek J, Lindstedt B-A et al. (2014) Exploiting the explosion of information associated with whole genome sequencing to tackle Shiga toxin-producing *Escherichia coli* (STEC) in global food production systems. International Journal of Food Microbiology 187:57-72.
Frey KG, Herrera-Galeano J, Redden CL, Luu TV, Servetas SL, Mateczun AJ, Mokashi VP, Bishop-Lilly KA (2014) Comparison of three next-generation sequencing platforms for metagenomic sequencing and identification of pathogens in blood. BMC Genomics 15:96-110.
Fricker AM, Podlesny D, Fricke WF (2019) What is new and relevant for sequencing-based microbiome research? A mini-review. J Adv Res 19:105-112.
Frohling A, Rademacher A, Rumpold B, Klocke M, Schluter O (2018) Screening of microbial communities associated with endive lettuce during postharvest processing on industrial scale. Heliyon 4:e00671.
Gaillard J-L, Berche P, Frehei C, Gouin E, Cossart P (1991) Entry of *L. monocytogenes* into cells is mediated by Internalin, a repeat protein reminiscent of surface antigens from Gram-positive cocci. Cell 65:1127-1141.
Galie S, Garcia-Gutierrez C, Miguelez EM, Villar CJ, Lombo F (2018) Biofilms in the food industry: health aspects and control methods. Frontiers in Microbiology 9:898.
Gandhi M, Chikindas ML (2007) *Listeria*: A foodborne pathogen that knows how to survive. International Journal of Food Microbiology 113:1-15.

Gandhi M, Golding S, Yaron S, Matthews KR (2001) Use of green fluorescent protein expressing *Salmonella S*tanley to investigate survival, spatial location, and control on alfalfa sprouts. Journal of Food Protection 64:1891-1898.

Ge C, Lee C, Nangle E, Li J, Gardner D, Kleinhenz M, Lee J (2014) Impact of phytopathogen infection and extreme weather stress on internalization of *Salmonella typhimurium* in lettuce. International Journal of Food Microbiology 168-169:24-31.

Gilchrist CA, Turner SD, Riley MF, Petri WA, Jr., Hewlett EL (2015) Whole-genome sequencing in outbreak analysis. Clinical Microbiology Reviews 28:541-563.

Gilmour MW, Graham M, Van Domselaar G, Tyler S, Kent H, Trout-Yakel KM, Larios O, Allen V, Lee B, Nadon C (2010) High-throughput genome sequencing of two *Listeria monocytogenes* clinical isolates during a large foodborne outbreak. BMC Genomics 11:120.

Golberg D, Kroupitski Y, Belausov E, Pinto R, Sela S (2011) *Salmonella typhimurium* internalization is variable in leafy vegetables and fresh herbs. International Journal of Food Microbiology 145:250-257.

Gormley FJ, Little CL, Rawal N, Gillespie IA, Lebaigue S, Adak GK (2010) A 17-year review of foodborne outbreaks: describing the continuing decline in England and Wales (1992–2008). Epidemiology and Infection 139:688-699.

Gourle H, Karlsson-Lindsjo O, Hayer J, Bongcam-Rudloff E (2018) Simulating Illumina Metagenomic Data with InSilicoSeq. Bioinformatics.

Greenberg HB, Estes MK (2009) Rotaviruses: from pathogenesis to vaccination. Gastroenterology 136:1939-1951.

Guo S, Tay MYF, Aung KT, Seow KLG, Ng LC, Purbojati RW, Drautz-Moses DI, Schuster SC, Schlundt J (2019) Phenotypic and genotypic characterization of antimicrobial resistant Escherichia coli isolated from ready-to-eat food in Singapore using disk diffusion, broth microdilution and whole genome sequencing methods. Food Control 99:89-97.

Guzzon R, Franciosi E, Larcher R (2014) A new resource from traditional wines: characterisation of the microbiota of “Vino Santo” grapes as a biocontrol agent against *Botrytis cinerea*. European Food Research and Technology 239:117-126.

Gwinn M, MacCannell D, Armstrong GL (2019) Next-Generation Sequencing of Infectious Pathogens. JAMA 321:893-894.

Gyles C, Boerlin P (2014) Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease. Veterinary Pathology 51:328-340.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M et al. (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols 8:1494-1512.

Hang J, Desai V, Zavaljevski N, Yang Y, Lin X, Satya R, Martinez LJ, Blaylock JM, Jarman RG, Thomas SJ et al. (2014) 16S rRNA gene pyrosequencing of reference and clinical samples and investigation of the temperature stability of microbiome profiles. Microbiome 2:31-46.

Hankin ME (1896) The bactericidal action of the waters of the Jamuna and Ganges rivers on Cholera microbes. Annales de l'Institut Pasteur 10:511-523.

Hanshew AS, Mason CJ, Raffa KF, Currie CR (2013) Minimization of chloroplast contamination in 16S rRNA gene pyrosequencing of insect herbivore bacterial communities. Journal of Microbiological Methods 95:149-155.

Harding CD, Shaw EG (1990) Antimicrobial activity of Leuconostoc gelidum against closely related species and Listeria monocytogenes. Journal of Applied Bacteriology 69:648-654.

Harwood VJ, Staley C, Badgley BD, Borges K, Korajkic A (2014) Microbial source tracking markers for detection of fecal contamination in environmental waters: Relationships between pathogens and human health outcomes. FEMS Microbiology Reviews 38:1-40.

Haynes E, Jimenez E, Pardo MA, Helyar SJ (2019) The future of NGS (Next Generation Sequencing) analysis in testing food authenticity. Food Control 101:134-143.

Health Protection Agency (now Public Health England) (2009) Guidelines for assessing the microbiological safety of ready to eat foods.

Heather JM, Chain B (2016) The sequence of sequencers: The history of sequencing DNA. Genomics 107:1-8.
Heaton JC, Jones K (2008a) Microbial contamination of fruit and vegetables and the behaviour of enteropathogens in the phyllosphere: a review. Journal of Applied Microbiology 104:613-626.
Heaton JC, Jones K (2008b) Microbial contamination of fruit and vegetables and the behaviour of enteropathogens in the phyllosphere: a review. J Appl Microbiol 104:613-626.
Henri C, Leekitcharoenphon P, Carleton HA, Radomski N, Kaas RS, Mariet JF, Felten A, Aarestrup FM, Gerner Smidt P, Roussel S et al. (2017) An Assessment of Different Genomic Approaches for Inferring Phylogeny of *Listeria monocytogenes*. Frontiers in Microbiology 8:2351.
Heravi FS, Zakrzewski M, Vickery K, Hu H (2020) Host DNA depletion efficiency of microbiome DNA enrichment methods in infected tissue samples. Journal of Microbiological Methods 170:105856.
Highmore CJ, Warner JC, Rothwell SD, Wilks SA, Keevil CW (2018) Viable-but-nonculturable *Listeria monocytogenes* and *Salmonella enterica* serovar Thompson induced by chlorine stress remain infectious. mBio 9:e00540-00518.
Holmes NA, Innocent TM, Heine D, Bassam MA, Worsley SF, Trottmann F, Patrick EH, Yu DW, Murrell JC, Schiott M et al. (2016) Genome analysis of two *Pseudonocardia* phylotypes associated with *Acromyrmex* leafcutter ants reveals their biosynthetic potential. Frontiers in Microbiology 7:2073.
Holzel CS, Tetens JL, Schwaiger K (2018) Unraveling the role of vegetables in spreading antimicrobial-resistant bacteria: A need for quantitative risk assessment. Foodborne Pathogens and Disease 15:671-688.
Huttenhower C (2019) Huttenhower Lab Galaxy Applications. http://huttenhower.sph.harvard.edu/galaxy/.
Hyeon J-Y, Li S, Mann DA, Zhang S, Li Z, Chen Y, Deng X (2017) Quasimetagenomics-based and real-time-sequencing-aided detection and subtyping of *Salmonella enterica* from food samples. Applied and Environmental Microbiology 84:e02340-02317.
Iglesias MB, Abadias M, Anguera M, Sabata J, Viñas I (2017) Antagonistic effect of probiotic bacteria against foodborne pathogens on fresh-cut pear. LWT - Food Science and Technology 81:243-249.
Illumina (2010) Illumina Sequencing Technology: Highest data accuracy, simple workflow, and a broad range of applications.
Illumina (2018a) MiSeq System Denature and Dilute Libraries Guide, Document # 15039740 v09.
Illumina (2018b) MiSeq System Guide, Document # 15027617 v04 Material # 20024228, Chapter 3 Sequencing.
Illumina (2018c) Nextera XT DNA Library Prep Kit 15031942 v03.
Illumina (2018d) ScriptSeq Complete Kit (Plant). https://www.illumina.com/products/scriptseq-plant.html. Accessed 16/05/2018
Issenhuth-Jeanjean S, Roggentin P, Mikoleit M, Guibourdenche M, de Pinna E, Nair S, Fields PI, Weill F-X (2014) Supplement 2008–2010 (no. 48) to the White–Kauffmann–Le Minor scheme. Research in Microbiology 165:526-530.
Jackson C, Stone B, Tyler H (2015) Emerging perspectives on the natural microbiome of fresh produce vegetables. Agriculture 5:170-187.
Jackson CR, Randolph KC, Osborn SL, Tyler HL (2013) Culture dependent and independent analysis of bacterial communities associated with commercial salad leaf vegetables. BMC Microbiology 13:274-286.
Jafari N, Behroozi R, Farajzadeh D, Farsi M, Akbari-Noghabi K (2014) Antibacterial activity of *Pseudonocardia* sp. JB05, a rare salty soil actinomycete against *Staphylococcus aureus*. BioMed Research International 2014:182945.
Jagadeesan B, Gerner-Smidt P, Allard MW, Leuillet S, Winkler A, Xiao Y, Chaffron S, Van Der Vossen J, Tang S, Katase M et al. (2019) The use of next generation sequencing for improving food safety: Translation into practice. Food Microbiology 79:96-115.

Jamuna M, Babusha ST, Jeevaratnam K (2005) Inhibitory efficacy of nisin and bacteriocins from *Lactobacillus* isolates against food spoilage and pathogenic organisms in model and food systems. Food Microbiology 22:449-454.

Janda JM, Abbott SL (2007) 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. Journal of Clinical Microbiology 45:2761-2764.

Janisiewicz WJ, Conway WS, Brown MW, Sapers GM, Fratamico P, Buchanan RL (1999) Fate of *Escherichia coli* O157:H7 on fresh-cut apple tissue and its potential for transmission by fruit flies. Applied and Environmental Microbiology 65:1-5.

Jarvis KG, White JR, Grim CJ, Ewing L, Ottesen AR, Beaubrun JJ-G, Pettengill JB, Brown E, Hanes DE (2015) Cilantro microbiome before and after nonselective pre-enrichment for *Salmonella* using 16S rRNA and metagenomic sequencing. BMC Microbiology 15.

Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, Lago BA, Dave BM, Pereira S, Sharma AN et al. (2017a) Gene *CRP*. https://card.mcmaster.ca/ontology/36657. Accessed August 2019

Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, Lago BA, Dave BM, Pereira S, Sharma AN et al. (2017b) Gene *H-NS*. https://card.mcmaster.ca/ontology/37020. Accessed August 2019

Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, Lago BA, Dave BM, Pereira S, Sharma AN et al. (2017c) Gene *mexB*. https://card.mcmaster.ca/ontology/36517. Accessed August 2019

Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, Lago BA, Dave BM, Pereira S, Sharma AN et al. (2017d) Gene *mexF*. https://card.mcmaster.ca/ontology/37184. Accessed August 2019

Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, Aarestrup FM (2014) Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of Verotoxigenic *Escherichia coli*. Journal of Clinical Microbiology 52:1501-1510.

Johnson BA, Anker H, Meleney FL (1945) Bacitracin: A new antibiotic produced by a member of the *B. subtilis* group. Science 102:376-377.

Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, Leopold SR, Hanson BM, Agresta HO, Gerstein M et al. (2019) Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun 10:5029.

Jones GS, D'Orazio SE (2013) *Listeria monocytogenes*: cultivation and laboratory maintenance. Current Protocols in Microbiology 31:9B 2 1-7.

Jones MB, Highlander SK, Anderson EL, Li W, Dayrit M, Klitgord N, Fabani MM, Seguritan V, Green J, Pride DT et al. (2015) Library preparation methodology can influence genomic and functional predictions in human microbiome research. Proceedings of the National Academy of Sciences 112:14024-14029.

Joshi NA, Fass JN (2011) Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle.

Kadam SR, den Besten HMW, van der Veen S, Zwietering MH, Moezelaar R, Abee T (2013) Diversity assessment of *Listeria monocytogenes* biofilm formation: impact of growth condition, serotype and strain origin. International Journal of Food Microbiology 165:259-264.

Kalani BS, Pournajaf A, Sedighi M, Bahador A, Irajian G, Valian F (2015) Genotypic characterization, invasion index and antimicrobial resistance pattern in *Listeria monocytogenes* strains isolated from clinical samples. Journal of Acute Disease 4:141-146.

Karki G (2017) Sanger's method of gene sequencing. https://www.onlinebiologynotes.com/sangers-method-gene-sequencing/. Accessed 24/10/2019

Kathariou S (2002) *Listeria monocytogenes* virulence and pathogenicity, a food safety perspective. Journal of Food Protection 65:1811-1829.

Kennedy K, Hall MW, Lynch MDJ, Moreno-Hagelsieb G, Neufeld JD (2014) Evaluating bias of Illumina-based bacterial 16S rRNA gene profiles. Applied and Environmental Microbiology 80:5717-5722.

Keys AL, Dailey RC, Hitchins AD, Smiley RD (2013) Postenrichment population differentials using buffered *Listeria* enrichment broth: Implications of the presence of *Listeria innocua* on *Listeria monocytogenes* in food test samples. Journal of Food Protection 76:1854-1862.
Kim HJ, Koo M, Hwang D, Choi JH, Kim SM, Oh S-W (2016) Contamination patterns and molecular typing of *Bacillus cereus* in fresh-cut vegetable salad processing. Applied Biological Chemistry 59:573-577.
Kurenbach B, Marjoshi D, Amabile-Cuevas CF, Ferguson GC, Godsoe W, Gibson P, Heinemann JA (2015) Sublethal exposure to commercial formulations of the herbicides dicamba, 2,4-dichlorophenoxyacetic acid, and glyphosate cause changes in antibiotic susceptibility in *Escherichia coli* and *Salmonella enterica* serovar Typhimurium. MBio 6.
Lakaye B, Dubus A, Lepage S, Groslambert S, Frere J-M (1999) When drug inactivation renders the target irrelevant to antibiotic resistance: a case story with b-lactams. Molecular Microbiology 31:89-101.
Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, Studholme DJ (2015) Assessing the performance of the Oxford Nanopore Technologies MinION. Biomolecular Detection and Quantification 3:1-8.
Lee J, Lim H, Jang H, Hwang B, Lee JH, Cho J, Lee JH, Bang D (2019) CRISPR-Cap: multiplexed double-stranded DNA enrichment based on the CRISPR system. Nucleic Acids Research 47:e1.
Lee RM, Lessler J, Lee RA, Rudolph KE, Reich NG, Perl TM, Cummings DAT (2013) Incubation periods of viral gastroenteritis: a systematic review. BMC Infectious Diseases 13:446.
Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J (2018) pyseer: a comprehensive tool for microbial pangenome-wide association studies. Bioinformatics 34:4310-4312.
Leonard SR, Mammel MK, Lacher DW, Elkins CA, Drake HL (2015) Application of metagenomic sequencing to food safety: Detection of Shiga toxin-producing *Escherichia coli* on fresh bagged spinach. Applied and Environmental Microbiology 81:8183-8191.
Leong D, Alvarez-Ordonez A, Jordan K (2014) Monitoring occurrence and persistence of *Listeria monocytogenes* in foods and food processing environments in the Republic of Ireland. Frontiers in Microbiology 5:436.
Lerner A, Matthias T, Aminov R (2017) Potential Effects of Horizontal Gene Exchange in the Human Gut. Frontiers in Immunology 8:1630.
Levy SB (2002) Active efflux, a common mechanism for biocide and antibiotic resistance. Journal of Applied Microbiology Symposium Supplement 92:65S-71S.
Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079.
Li L, Mendis N, Trigui H, Oliver JD, Faucher SP (2014) The importance of the viable but non-culturable state in human bacterial pathogens. Frontiers in Microbiology 5:258.
Life Technologies (2015a) Qubit® dsDNA HS Assay Kits MAN0002326 | MP32851. B.0 edn.
Life Technologies (2015b) Qubit® RNA HS Assay Kits MAN0002327 | MP32852. vol A.0.
Lima L, Marchet C, Caboche S, Da Silva C, Istace B, Aury JM, Touzet H, Chikhi R (2019) Comparative assessment of long-read error correction software applied to Nanopore RNA-sequencing data. Brief Bioinform.
Little CL, Gillespie IA (2008) Prepared salads and public health. Journal of Applied Microbiology 105:1729-1743.
Lu J, Breitwieser FP, Thielen P, Salzberg SL (2017) Bracken: estimating species abundance in metagenomics data. PeerJ Computer Science 3:https://doi.org/10.7717/peerj-cs.7104.
Maćkiw E, Modzelewska M, Mąka Ł, Ścieżyńska H, Pawłowska K, Postupolski J, Korsak D (2016) Antimicrobial resistance profiles of *Listeria monocytogenes* isolated from ready-to-eat products in Poland in 2007–2011. Food Control 59:7-11.
Magajna BA, Schraft H (2015) *Campylobacter jejuni* biofilm cells become viable but non-culturable (VBNC) in low nutrient conditions at 4 °C more quickly than their planktonic counterparts. Food Control 50:45-50.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380.

Marti R, Scott A, Tien Y-C, Murray R, Sabourin L, Zhang Y, Topp E (2013) Impact of Manure Fertilization on the Abundance of Antibiotic-Resistant Bacteria and Frequency of Detection of Antibiotic Resistance Genes in Soil and on Vegetables at Harvest. Applied and Environmental Microbiology 79:5701-5709.

Martínez JL, Baquero F, Andersson DI (2007) Predicting antibiotic resistance. Nature Reviews Microbiology 5:958-965.

Martinis ECPD, Franco BDGM (1998) Inhibition of *Listeria monocytogenes* in a pork product by a *Lactobacillus sake* strain. International Journal of Food Microbiology 42:119-126.

Matias BG, Pinto PSA, Cossi MVC, Silva A, Vanetti MCD, Nero LA (2010) Evaluation of a polymerase chain reaction protocol for the detection of *Salmonella* species directly from superficial samples of chicken carcasses and preenrichment broth. Poultry Science 89:1524-1529.

Maurice CF, Haiser HJ, Turnbaugh PJ (2013) Xenobiotics shape the physiology and gene expression of the active human gut microbiome. Cell 152:39-50.

Maury MM, Chenal-Francisque V, Bracq-Dieye H, Han L, Leclercq A, Vales G, Moura A, Gouin E, Scortti M, Disson O et al. (2017) Spontaneous loss of virulence in natural populations of *Listeria monocytogenes*. Infection and Immunity 85:e00541-00517.

Maury MM, Tsai Y-H, Charlier C, Touchon M, Chenal-Francisque V, Leclercq A, Criscuolo A, Gaultier C, Roussel S, Brisabois A et al. (2016) Uncovering *Listeria monocytogenes* hypervirulence by harnessing its biodiversity. Nature Genetics 48:308-313.

McLaughlin HP, Casey PG, Cotter J, Gahan CGM, Hill C (2011) Factors affecting survival of *Listeria monocytogenes* and *Listeria innocua* in soil samples. Archives of Microbiology 193:775-785.

Mendes R, Garbeva P, Raaijmakers JM (2013) The rhizosphere microbiome: Significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. FEMS Microbiology Reviews 37:634-663.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A et al. (2008) The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9:386-394.

Middlemiss JK, Poole K (2004) Differential impact of *MexB* mutations on substrate selectivity of the MexAB-OprM multidrug efflux pump of *Pseudomonas aeruginosa*. Journal of Bacteriology 186:1258-1269.

Mohammadpour H, Berizi E, Hosseinzadeh S, Majlesi M, Zare M (2018) The prevalence of Campylobacter spp. in vegetables, fruits, and fresh produce: a systematic review and meta-analysis. Gut Pathogens 10:41.

Monaghan J, Hutchinson M (2010) Monitoring microbial food safety of fresh produce.

Monaghan JM, Thomas DJI, Goodburn K, Hutchinson ML (2008) A review of the published literature describing foodborne illness outbreaks associated with ready to eat fresh produce and an overview of current UK fresh produce farming practices (trans: B17007 FSAP).

Nachamkin I, Allos BM, Ho T (1998) *Campylobacter* species and Guillain-Barre syndrome. Clinical Microbiology Reviews 11:555-567.

Nakamura T, Yamada KD, Tomii K, Katoh K (2018) Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics 34:2490-2492.

Neuert S, Nair S, Day MR, Doumith M, Ashton PM, Mellor KC, Jenkins C, Hopkins KL, Woodford N, de Pinna E et al. (2018) Prediction of Phenotypic Antimicrobial Resistance Profiles From Whole Genome Sequences of Non-typhoidal Salmonella enterica. Frontiers in Microbiology 9:592.

New England Biolabs (2017) NEBNext Ultra II RNA Library Prep Kit for Illumina, Version 1.0, 5/17.

New England Biolabs (2018) NEBNext® Poly(A) mRNA Magnetic Isolation Module (NEB #E7490) Version 6.0 4/18.

Newton RJ, Bootsma MJ, Morrison HG, Sogin ML, McLellan SL (2013) A microbial signature approach to identify fecal pollution in the waters off an urbanized coast of Lake Michigan. Microbial Ecology 65:1011-1023.

NHS (2014) *Salmonella* infection. http://www.nhs.uk/Conditions/salmonella-infection/Pages/Introduction.aspx. Accessed 06/01/2017
NHS (2015a) *Escherichia coli* (*E. coli*) O157. http://www.nhs.uk/conditions/Escherichia-Coli-O157/Pages/Introduction.aspx. Accessed 06/01/2017
NHS (2015b) Food poisoning - Causes http://www.nhs.uk/Conditions/Food-poisoning/Pages/Causes.aspx. Accessed 06/01/2017
NHS (2015c) Listeriosis - Symptoms. http://www.nhs.uk/Conditions/Listeriosis/Pages/Symptoms.aspx. Accessed 06/01/2017
NHS (2016a) Diarrhoea and vomiting (gastroenteritis). http://www.nhs.uk/Conditions/Gastroenteritis/Pages/Introduction.aspx. Accessed 06/01/2017
NHS (2016b) Hepatitis A. http://www.nhs.uk/conditions/Hepatitis-A/Pages/Introduction.aspx. Accessed 06/01/2017
NHS (2017) Rotavirus vaccine. https://www.nhs.uk/conditions/vaccinations/rotavirus-vaccine/. Accessed 12/06/2019
Nishibori T, Cooray K, Xiong H, Kawamura I, Fujita M, Mitsuyama M (1995) Correlation between the presence of virulence-associated genes as determined by PCR and actual virulence to mice in various strains of *Listeria* spp. . Medical Microbiology and Immunology 39:343-349.
Nishino K, Senda Y, Yamaguchi A (2008) *CRP* Regulator Modulates Multidrug Resistance of *Escherichia coli* by Repressing the *mdtEF* Multidrug Efflux Genes. The Journal of Antibiotics 61:120-127.
Nishino K, Yamaguchi A (2004) Role of histone-like protein H-NS in multidrug resistance of *Escherichia coli*. Journal of Bacteriology 186:1423-1429.
Noll M, Kleta S, Al Dahouk S (2018) Antibiotic susceptibility of 259 *Listeria monocytogenes* strains isolated from food, food-processing plants and human samples in Germany. Journal of Infection and Public Health 11:572-577.
Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, Prjibelsky A, Pyshkin A, Sirotkin A, Sirotkin Y et al. Assembling genomes and mini-metagenomes from highly chimeric reads. In, Berlin, Heidelberg, 2013. Research in Computational Molecular Biology. Springer Berlin Heidelberg, pp 158-170.
O'Brien SJ, Larose TL, Adak GK, Evans MR, Tam CC (2016) Modelling study to estimate the health burden of foodborne diseases: cases, general practice consultations and hospitalisations in the UK, 2009. BMJ Open 6:e011119.
Obaidat MM, Bani Salman AE, Lafi SQ, Al-Abboodi AR (2015) Characterization of *Listeria monocytogenes* from three countries and antibiotic resistance differences among countries and *Listeria monocytogenes* serogroups. Letters in Applied Microbiology 60:609-614.
Ogunade IM, Kim DH, Jiang Y, Weinberg ZG, Jeong KC, Adesogan AT (2016) Control of *Escherichia coli* O157:H7 in contaminated alfalfa silage: Effects of silage additives. Journal of Dairy Science 99:4427-4436.
Olaimat AN, Al-Holy MA, Shahbaz HM, Al-Nabulsi AA, Abu Ghoush MH, Osaili TM, Ayyash MM, Holley RA (2018) Emergence of antibiotic resistance in *Listeria monocytogenes* isolated from food products: A comprehensive review. Comprehensive Reviews in Food Science and Food Safety 17:1277-1292.
Olaimat AN, Holley RA (2012) Factors influencing the microbial safety of fresh produce: A review. Food Microbiology 32:1-19.
Oliver JD (2010) Recent findings on the viable but nonculturable state in pathogenic bacteria. FEMS Microbiology Reviews 34:415-425.
Ölmez H, Temur SD (2010) Effects of different sanitizing treatments on biofilms and attachment of *Escherichia coli* and *Listeria monocytogenes* on green leaf lettuce. LWT - Food Science and Technology 43:964-970.
Olsen I (2015) Biofilm-specific antibiotic tolerance and resistance. European Journal of Clinical Microbiology & Infectious Diseases 34:877-886.

Oravcová K, Trnčíková T, Kuchta T, Kaclíková E (2007) Limitation in the detection of *Listeria monocytogenes* in food in the presence of competing *Listeria innocua*. Journal of Applied Microbiology 104:429-437.

Orlofsky E, Bernstein N, Sacks M, Vonshak A, Benami M, Kundu A, Maki M, Smith W, Wuertz S, Shapiro K et al. (2016) Comparable levels of microbial contamination in soil and on tomato crops after drip irrigation with treated wastewater or potable water. Agriculture, Ecosystems & Environment 215:140-150.

Ottesen AR, Gonzalez Pena A, White JR, Pettengill JB, Li C, Allard S, Rideout S, Allard M, Hill T, Evans P et al. (2013) Baseline survey of the anatomical microbial ecology of an important food plant: *Solanum lycopersicum* (tomato). BMC Microbiology 13:114-125.

Pacific Biosciences (2015) Revolutionize Genomics with SMRT® Sequencing. Single Molecule, Real-Time Technology.

Painset A, Bjorkman JT, Kiil K, Guillier L, Mariet JF, Felix B, Amar C, Rotariu O, Roussel S, Perez-Reche F et al. (2019) LiSEQ - whole-genome sequencing of a cross-sectional survey of *Listeria monocytogenes* in ready-to-eat foods and human clinical cases in Europe. Microbial Genomics 5:e000257.

Pearman WS, Freed NE, Silander OK (2019) The advantages and disadvantages of short- and long-read metagenomics to infer bacterial and eukaryotic community composition. bioRxiv.

Pedroso A, Hurley-Bacon A, Zedek A, Kwan T, Jordan A, Avellaneda G, Hofacre C, Oakley B, Collett S, Maurer J et al. (2013) Can probiotics improve the environmental microbiome and resistome of commercial poultry production? International Journal of Environmental Research and Public Health 10:4534-4559.

Pintó RM, Costafreda MI, Pérez-Rodriguez FJ, D'Andrea L, Bosch A (2010) Hepatitis A Virus: State of the Art. Food and Environmental Virology 2:127-135.

Pirone-Davies C, Chen Y, Pightling A, Ryan G, Wang Y, Yao K, Hoffmann M, Allard MW (2018) Genes significantly associated with lineage II food isolates of *Listeria monocytogenes*. BMC Genomics 19:708-719.

Pollock J, Glendinning L, Wisedchanwet T, Watson M (2018) The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies. Applied and Environmental Microbiology 84.

Poole K (2002) Mechanisms of bacterial biocide and antibiotic resistance. Journal of Applied Microbiology Symposium Supplement 92:55S-64S.

Price RJ, Lee JS (1970) Inhibition of *Pseudomonas* species by hydrogen peroxide producing *Lactobacilli*. Journal of Milk and Food Technology 33:13-18.

Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK (2016) Genomics and taxonomy in diagnostics for food security: soft-rotting *Enterobacterial* plant pathogens. Analytical Methods 8:12-24.

Public Health England (2019) *Listeria* cases being investigated https://www.gov.uk/government/news/listeria-cases-being-investigated. Accessed 18/07/2019

Puga CH, Dahdouh E, SanJose C, Orgaz B (2018) *Listeria monocytogenes c*olonizes *Pseudomonas fluorescens b*iofilms and induces matrix over-production. Frontiers in Microbiology 9:1706.

Qiagen (2005) AllPrep DNA/RNA Mini Handbook 11/2005.

Qiagen (2006) DNeasy® Blood & Tissue Handbook 07/2006.

Qiagen (2012a) mericon Pathogen Detection Handbook 07/2012

Qiagen (2012b) RNeasy® Mini Handbook 06/2012.50-53, 67, 68.

Qiime2docs (2017a) Importing data. https://docs.qiime2.org/2017.12/tutorials/importing/. Accessed 25/01/2018

Qiime2docs (2017b) "Moving Pictures" tutorial. https://docs.qiime2.org/2017.12/tutorials/moving-pictures/. Accessed 25/01/2018

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13:341-354.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research 41:D590-596.
Quesada E, Ventosa A, Rodriguez-Valera F, Megias L, Ramos-Cormenzana A (1983) Numerical taxonomy of moderately halophilic Gram-negative bacteria from hypersaline soils. Journal of General Microbiology 129:2649-2657.
Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-842.
Reyneke B, Ndlovu T, Khan S, Khan W (2017) Comparison of EMA-, PMA- and DNase qPCR for the determination of microbial cell viability. Applied Microbiology and Biotechnology 101:7371-7383.
Rico D, Martín-Diana AB, Barat JM, Barry-Ryan C (2007) Extending and measuring the quality of fresh-cut fruit and vegetables: a review. Trends in Food Science & Technology 18:373-386.
Robilotti E, Deresinski S, Pinsky BA (2015) Norovirus. Clinical Microbiology Reviews 28:134-164.
Rodriguez-Lopez P, Rodriguez-Herrera JJ, Vazquez-Sanchez D, Lopez Cabo M (2018) Current knowledge on *Listeria monocytogenes* biofilms in food-related environments: incidence, resistance to biocides, ecology and biocontrol. Foods 7:85-104.
Rolain J-M (2013) Food and human gut as reservoirs of transferable antibiotic resistance encoding genes. Frontiers in Microbiology 4:173.
Rosimin AA, Kim M-J, Joo I-S, Suh S-H, Kim K-S (2016) Simultaneous detection of pathogenic *Listeria* including atypical *Listeria innocua* in vegetables by a quadruplex PCR method. LWT - Food Science and Technology 69:601-607.
Russell AD (2003) Similarities and differences in the responses of microorganisms to biocides. Journal of Antimicrobial Chemotherapy 52:750-763.
Sallach JB, Zhang Y, Hodges L, Snow D, Li X, Bartelt-Hunt S (2015) Concomitant uptake of antimicrobials and *Salmonella* in soil and into lettuce following wastewater irrigation. Environmental Pollution 197:269-277.
Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the USA 74:5463-5467.
Scallan E, Griffin PM, Angulo FJ, Tauxe RV, Hoekstra RM (2011) Foodborne illness acquired in the United States - Unspecified agents. Emerging Infectious Diseases 17:16-22.
Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ et al. (2009) Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Applied and Environmental Microbiology 75:7537-7541.
Seemann T (2018) ABRicate Github. https://github.com/tseemann/abricate. Accessed October
Seemann T, Silva AGd, Bulach DM, Schultz MB, Kwong JC, Howden BP (2018) Nullarbor Github. https://github.com/tseemann/nullarbor. Accessed December
Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett W, Huttenhower C (2011) Metagenomic biomarker discovery and explanation. Genome Biology 12:R60.
Sela Saldinger S, Manulis-Sasson S (2015) What else can we do to mitigate contamination of fresh produce by foodborne pathogens? Microbial Biotechnology 8:29-31.
Shen T, Pajaro-Van de Stadt SH, Yeat NC, Lin JC (2015) Clinical applications of next generation sequencing in cancer: from panels, to exomes, to genomes. Front Genet 6:215.
Shih SY, Bose N, Goncalves ABR, Erlich HA, Calloway CD (2018) Applications of Probe Capture Enrichment Next Generation Sequencing for Whole Mitochondrial Genome and 426 Nuclear SNPs for Forensically Challenging Samples. Genes (Basel) 9.
Signoretto C, Lleo MDM, Tafi MC, Canepari P (2000) Cell wall chemical composition of *Enterococcus faecalis* in the viable but nonculturable state. Applied and Environmental Microbiology 66:1953-1959.
Silva M, Machado MP, Silva DN, Rossi M, Moran-Gilad J, Santos S, Ramirez M, Carrico JA (2018) chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. Microbial Genomics 4:e000166.

Smith A, Hearn J, Taylor C, Wheelhouse N, Kaczmarek M, Moorhouse E, Singleton I (2019) *Listeria monocytogenes* isolates from ready to eat plant produce are diverse and have virulence potential. International Journal of Food Microbiology 299:23-32.

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SBH, Hood LE (1986) Fluorescence detection in automated DNA sequence analysis. Nature 321:674-679.

Song L, Xie K (2020) Engineering CRISPR/Cas9 to mitigate abundant host contamination for 16S rRNA gene-based amplicon sequencing. Microbiome 8:80.

Spann CT, Taylor SC, Weinberg JM (2004) Topical antimicrobial agents in dermatology. Disease-a-Month 50:407-421.

Sun H, Zu Y (2015) A Highlight of Recent Advances in Aptamer Technology and Its Application. Molecules 20:11959-11980.

Szmolka A, Nagy B (2013) Multidrug resistant commensal *Escherichia coli* in animals and its impact for public health. Frontiers in Microbiology 4:258.

Tam CC, Larose T, O'Brien SJ (2014) Costed extension to the Second Study of Infectious Intestinal Disease in the Community: Identifying the proportion of foodborne disease in the UK and attributing foodborne disease by food commodity. IID2 extension report. UK Food Standards Agency.

Tatsika S, Karamanoli K, Karayanni H, Genitsaris S (2019) Metagenomic characterization of bacterial communities on ready-to-eat vegetables and effects of household washing on their diversity and composition. Pathogens 8.

The European Committee on Antimicrobial Susceptibility Testing (2017) EUCAST Disk Diffusion Method for Antimicrobial Susceptibility Testing - Version 6.0 (January 2017).

The European Committee on Antimicrobial Susceptibility Testing (2018) Breakpoint tables for interpretation of MICs and zone diameters. Version 8.1.

Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G et al. (2017) A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457-463.

Tomás-Callejas A, López-Velasco G, Camacho AB, Artés F, Artés-Hernández F, Suslow TV (2011) Survival and distribution of *Escherichia coli* on diverse fresh-cut baby leafy greens under preharvest through postharvest conditions. International Journal of Food Microbiology 151:216-222.

Torchia MT, Austin DC, Kunkel ST, Dwyer KW, Moschetti WE (2019) Next-Generation Sequencing vs Culture-Based Methods for Diagnosing Periprosthetic Joint Infection After Total Knee Arthroplasty: A Cost-Effectiveness Analysis. Journal of Arthroplasty 34:1333-1341.

Trček J, Mahnič A, Rupnik M (2016) Diversity of the microbiota involved in wine and organic apple cider submerged vinegar production as revealed by DHPLC analysis and next-generation sequencing. International Journal of Food Microbiology 223:57-62.

Trias R, Badosa E, Montesinos E, Baneras L (2008a) Bioprotective *Leuconostoc* strains against *Listeria monocytogenes* in fresh fruits and vegetables. International Journal of Food Microbiology 127:91-98.

Trias R, Bañeras R, Badosa E, Montesinos E (2008b) Bioprotection of golden delicious apples and iceberg lettuce against foodborne bacterial pathogens by lactic acid bacteria. International Journal of Food Microbiology 123:50-60.

Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods 12:902-903.

U.S. Food and Drug Administration (2014) FDA Approval of *Listeria*-specific Bacteriophage Preparation on Ready-to-Eat (RTE) Meat and Poultry Products. https://www.fda.gov/food/ingredientspackaginglabeling/ucm083572.htm. Accessed 06/06/2017

U.S. Food and Drug Administration (2019) GenomeTrackr Fast Facts. https://www.fda.gov/food/whole-genome-sequencing-wgs-program/genometrakr-fast-facts. Accessed 03/10/2019

Van der Poel WHM, Dalton HR, Johne R, Pavio N, Bouwknegt M, Wu T, Cook N, Meng XJ (2018) Knowledge gaps and research priorities in the prevention and control of hepatitis E virus infection. Transboundary and Emerging Diseases 65 Suppl 1:22-29.
van Elsas JD, Chiurazzi M, Mallon CA, Elhottova D, Kristufek V, Salles JF (2012) Microbial diversity determines the invasion of soil by a bacterial pathogen. Proceedings of the National Academy of Sciences 109:1159-1164.
Verhoeff-Bakkenes L, Jansen HA, in 't Veld PH, Beumer RR, Zwietering MH, van Leusden FM (2011) Consumption of raw vegetables and fruits: a risk factor for *Campylobacter* infections. International Journal of Food Microbiology 144:406-412.
Vihavainen EJ, Murros AE, Bjorkroth KJ (2008) *Leuconostoc* spoilage of vacuum-packaged vegetable sausages. Journal of Food Protection 71:2312-2315.
Wales A, Davies R (2015) Co-selection of resistance to antibiotics, biocides and heavy metals, and its relevance to foodborne pathogens. Antibiotics 4:567-604.
Wall DH, Nielsen UN, Six J (2015) Soil biodiversity and human health. Nature 528:69-76.
Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10:57-63.
Warriner K, Namvar A (2010) The tricks learnt by human enteric pathogens from phytopathogens to persist within the plant environment. Current Opinion in Biotechnology 21:131-136.
Warriner K, Spaniolas S, Dickinson M, Wright C, Waites WM (2003) Internalization of bioluminescent *Escherichia coli* and *Salmonella Montevideo* in growing bean sprouts. Journal of Applied Microbiology 95:719-727.
Wells JM, Butterfield JE (1997) *Salmonella* contamination associated with bacterial soft rot of fresh fruits and vegetables in the marketplace. Plant Disease 81:867-872.
Wetterstrand KA (2019) DNA sequencing costs: Data from the NHGRI Genome Sequencing Program (GSP). https://www.genome.gov/sequencingcostsdata/. Accessed 24/10/2019
Wilson A, Gray J, Chandry PS, Fox EM (2018) Phenotypic and genotypic analysis of antimicrobial resistance among *Listeria monocytogenes i*solated from Australian food production chains. Genes 9:10.3390/genes9020080.
Wonderling LD, Wilkinson BJ, Bayles DO (2004) The *htrA* (*degP*) gene of *Listeria monocytogenes* 10403S is essential for optimal growth under stress conditions. Applied and Environmental Microbiology 70:1935-1943.
Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments Genome Biology:https://doi.org/10.1186/gb-2014-1115-1183-r1146.
Wood JD, Bezanson GS, Gordon RJ, Jamieson R (2010) Population dynamics of *Escherichia coli* inoculated by irrigation into the phyllosphere of spinach grown under commercial production conditions. International Journal of Food Microbiology 143:198-204.
World Health Organisation (2015a) Global Action Plan on Antimicrobial Resistance.
World Health Organisation (2015b) WHO estimates of the global burden of foodborne diseases: foodborne disease burden epidemiology reference group 2007-2015.
World Health Organisation (2016a) *Campylobacter* factsheet. http://www.who.int/mediacentre/factsheets/fs255/en/. Accessed 06/01/2017
World Health Organisation (2016b) *E. coli*. http://www.who.int/mediacentre/factsheets/fs125/en/. Accessed 06/01/2017
World Health Organisation (2016c) *Salmonella* (non-typhoidal) factsheet. http://www.who.int/mediacentre/factsheets/fs139/en/. Accessed 06/01/2017
Wylezich C, Papa A, Beer M, Hoper D (2018) A Versatile Sample Processing Workflow for Metagenomic Pathogen Detection. Scientific Reports 8:13108.
Yi L, Su G, Hu G, Peng Q (2017) Diversity study of microbial community in bacon using metagenomic analysis. Journal of Food Safety:https://doi.org/10.1111/jfs.12334.
Yurgel SN, Abbey L, Loomer N, Gillis-Madden R, Mammoliti M (2018) Microbial communities associated with storage onion. Phytobiomes Journal 2:35-41.

Zhang S, Wu Q, Zhang J, Lai Z, Zhu X (2016a) Prevalence, genetic diversity, and antibiotic resistance of enterotoxigenic *Escherichia coli* in retail ready-to-eat foods in China. Food Control 68:236-243.

Zhang Y, Sallach JB, Hodges L, Snow DD, Bartelt-Hunt SL, Eskridge KM, Li X (2016b) Effects of soil texture and drought stress on the uptake of antibiotics and the internalization of *Salmonella* in lettuce following wastewater irrigation. Environmental Pollution 208:523-531.

Zhu B, Chen Q, Chen S, Zhu Y-G (2016) Does organically produced lettuce harbor higher abundance of antibiotic resistance genes than conventionally produced? Environment International:http://dx.doi.org/10.1016/j.envint.2016.1011.1001.

Zwietering MH, Jacxsens L, Membré J-M, Nauta M, Peterz M (2016) Relevance of microbial finished product testing in food safety management. Food Control 60:31-43.

Appendices

Appendix A. Agencourt AMPure XP bead clean up protocol for post indexing samples

Preparation:

- Allow AMPure XP beads to come to room temperature and vortex thoroughly before use
- Make up 80% ethanol for washing your samples within an hour of performing the protocol

Procedure:

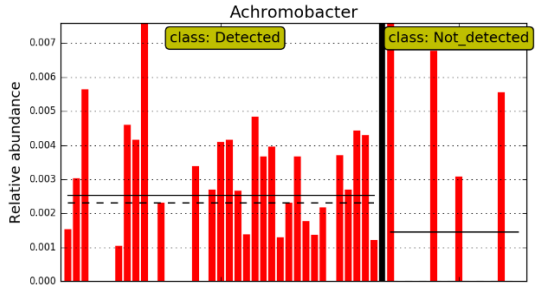
1. Add indicated quantity of beads to the sample and mix gently by pipetting up and down 10 times.
2. Incubate for 5 mins at room temperature
3. Place plate on magnetic stand and leave until the supernatant has cleared
4. With the plate on the stand, remove the supernatant and discard, being careful not to disturb the pellet
5. With the plate on the stand, wash the beads with freshly prepared 80% Ethanol twice as outlined below:
    a. Add 200 µl of 80% ethanol to each sample well
    b. Incubate for 30 s at room temperature
    c. Remove and discard the supernatant, being careful not to disturb the pellet
    d. Repeat for a second wash
6. Spin plate briefly to collect residual ethanol and remove using a p10 pipette with fine tip
7. Leave on magnetic stand to air dry at room temperature for 5-10 mins, or until completely dry
8. Remove the plate from the stand and add 47.5 µl MBGW to the samples, ensuring the pellet is fully resuspended,
9. Incubate for 5 mins at room temperature
10. Place back onto the magnetic stand until the supernatant has cleared
11. Transfer 45 µl of supernatant to a new well, ensuring no beads are transferred

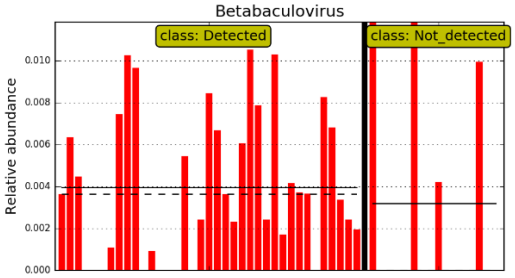
Appendix B. Mock microbiome dilutions series – proportion of each microbiome member per sample

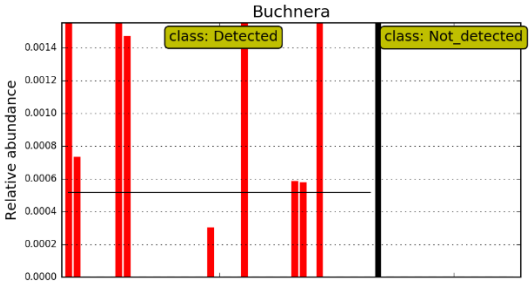
Accession Number	Name	Dilution series (proportion of 1)						
		100,000	10,000	1,000	100	10	1	0
NC 002944.2	*Mycobacterium avium* subsp. paratuberculosis str. k10	0.015	0.02	0.02	0.02	0.02	0.02	0.02
NC 002947.4	*Pseudomonas putida* KT2440 chromosome	0.02	0.04	0.04	0.04	0.04	0.04	0.04
NC 004722.1	*Bacillus cereus* ATCC 14579 chromosome	0.02	0.03	0.03	0.03	0.03	0.03	0.03
NC 007578.1	*Lactuca sativa chloroplast*	0.8	0.85	0.85	0.85	0.85	0.85	0.85
NC 016830.1	*Pseudomonas fluorescens* F113	0.015	0.02	0.02	0.02	0.02	0.02	0.02
NZ CP010519.1	*Streptomyces albus* strain DSM 41398	0.01	0.01	0.01	0.01	0.01	0.01	0.01
NZ CP011007.1	*Bacillus pumilus* strain SH-B9	0.01	0.01	0.019	0.0199	0.01999	0.019999	0.02
NZ LT700188.1	*Negativicoccus massiliensis* strain Marseille-P2082 genome assembly	0.01	0.01	0.01	0.01	0.01	0.01	0.01
NC 003198.1	*Salmonella enterica* subsp. enterica serovar Typhi str. CT18	0.1	0.01	0.001	0.0001	0.00001	0.000001	0

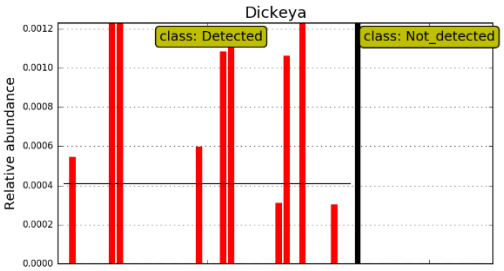
Appendix C. Differential Features from Lefse

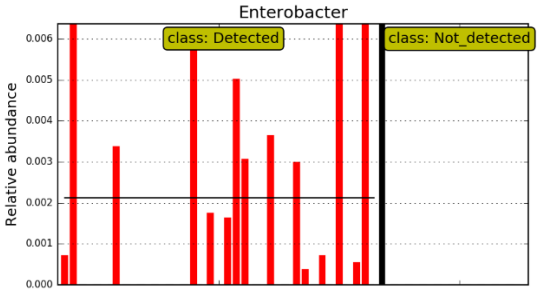
i) Differential features of association with positive or negative microbiology results for Enterobacteriaceae

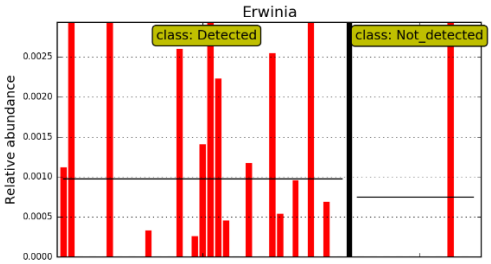


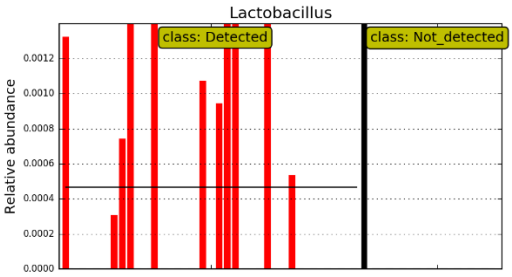


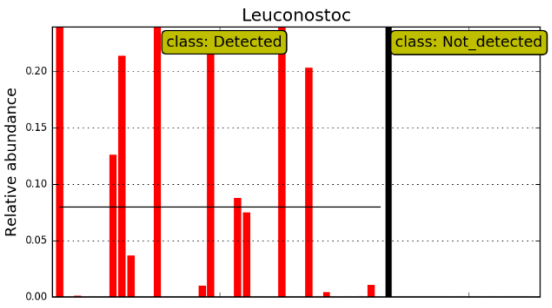


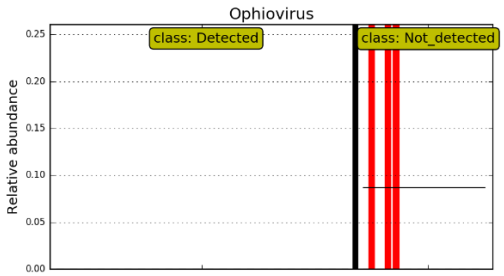


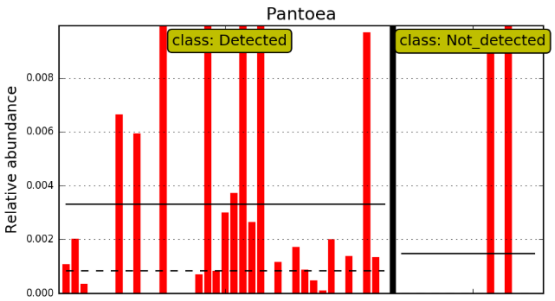


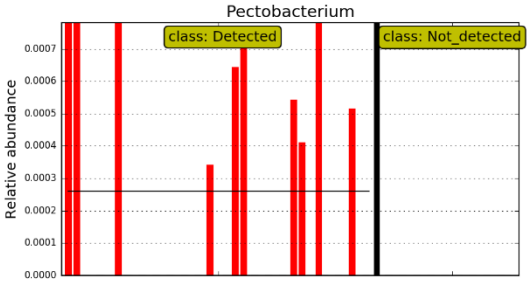


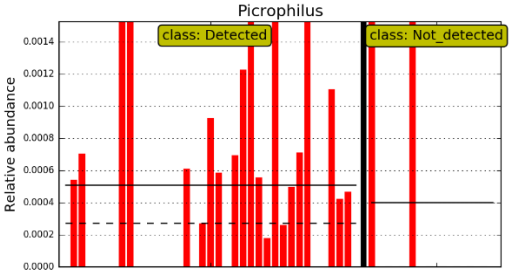


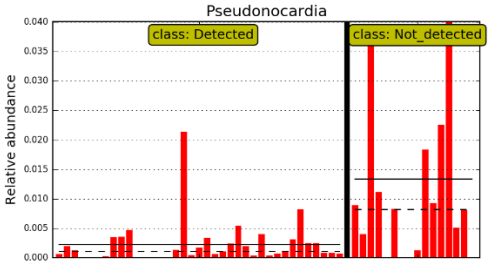


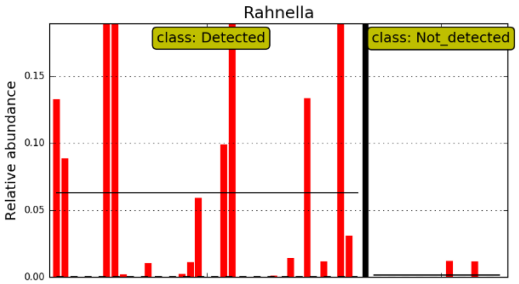

Ophiovirus
class: Detected
class: Not_detected
Relative abundance

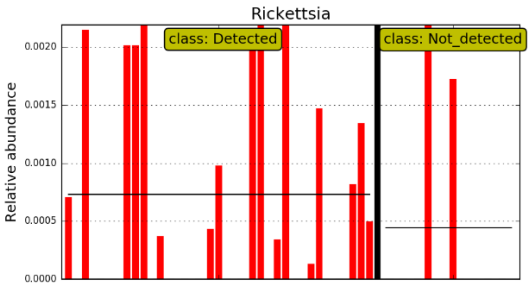

Pantoea
class: Detected
class: Not_detected
Relative abundance

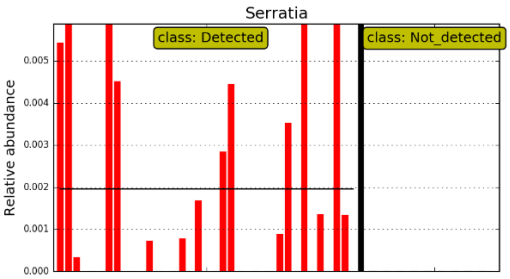

Pectobacterium
class: Detected
class: Not_detected
Relative abundance

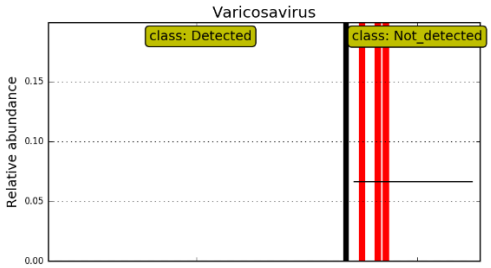

Picrophilus
class: Detected
class: Not_detected
Relative abundance

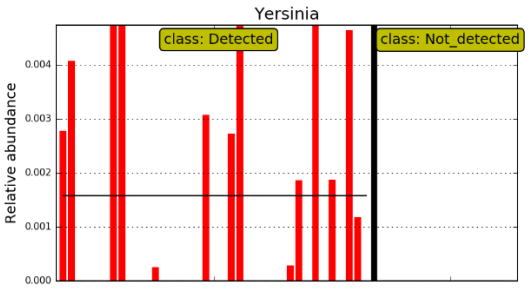

Pseudonocardia
class: Detected
class: Not_detected
Relative abundance


Rahnella
class: Detected
class: Not_detected
Relative abundance

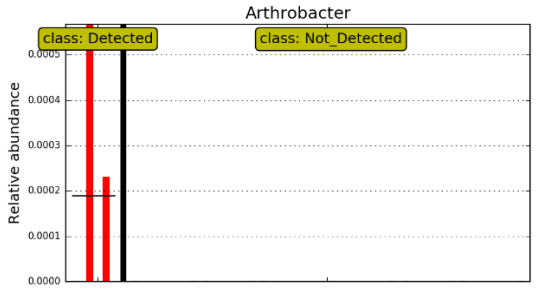

Rickettsia
class: Detected
class: Not_detected
Relative abundance

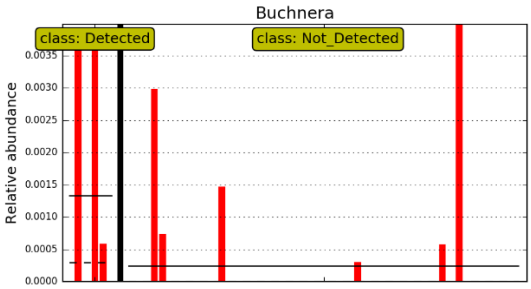

Serratia
class: Detected
class: Not_detected
Relative abundance

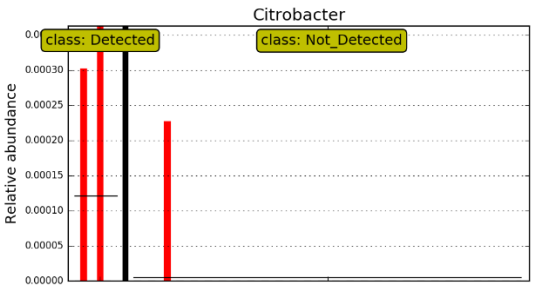

Varicosavirus
class: Detected
class: Not_detected
Relative abundance

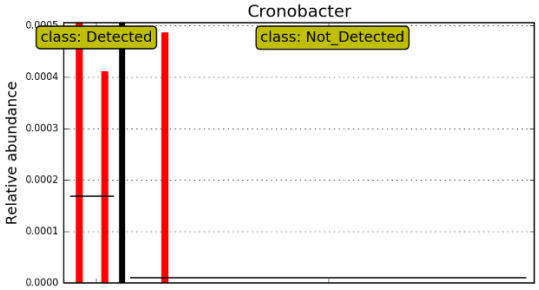

Yersinia
class: Detected
class: Not_detected
Relative abundance

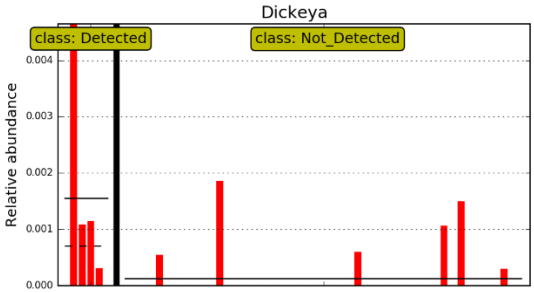
ii) Differential features of association with positive or negative microbiology results for Listeria spp.

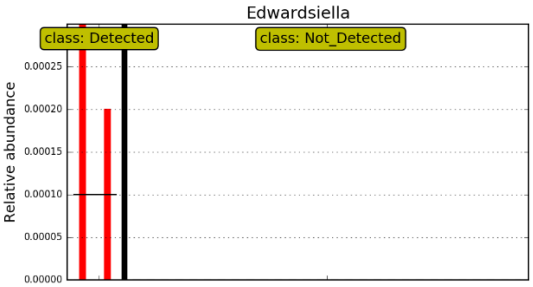


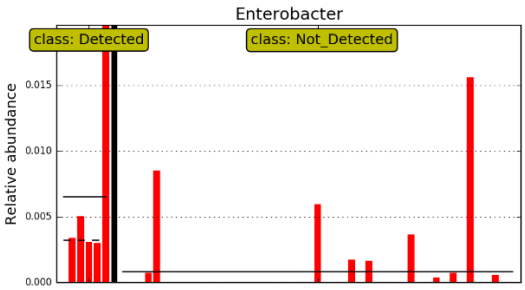


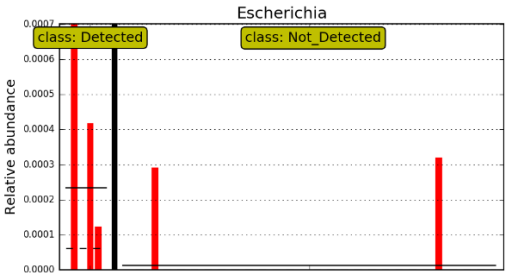


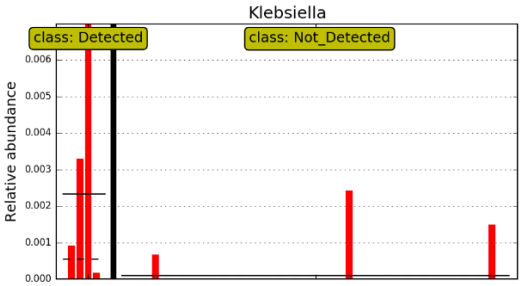


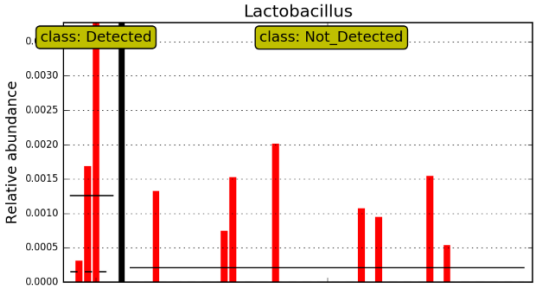


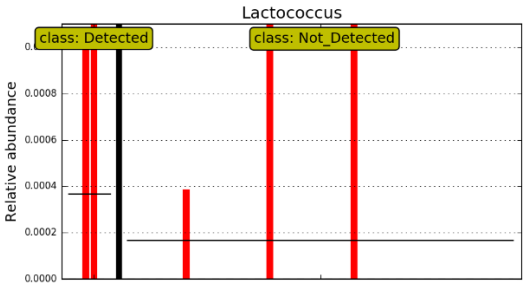


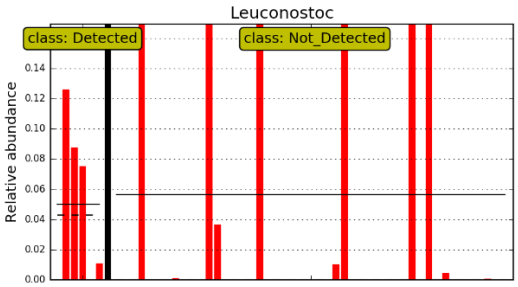


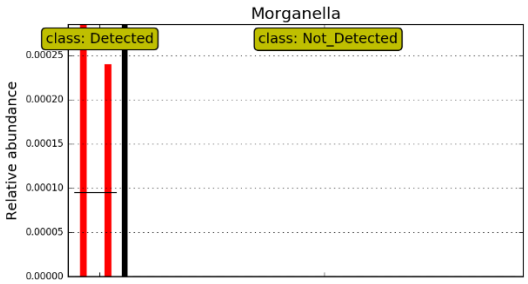


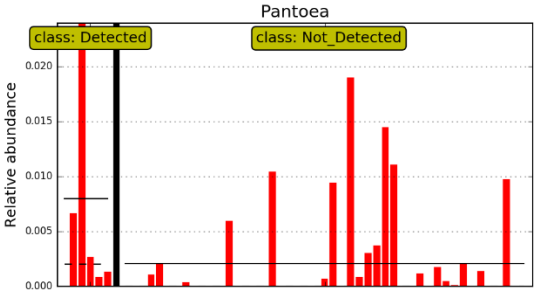


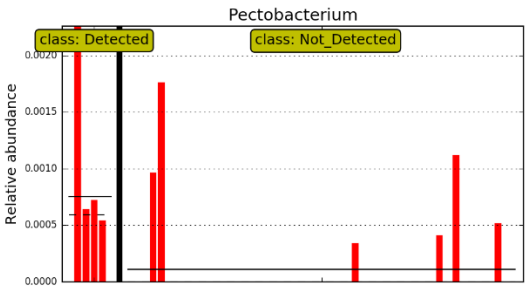


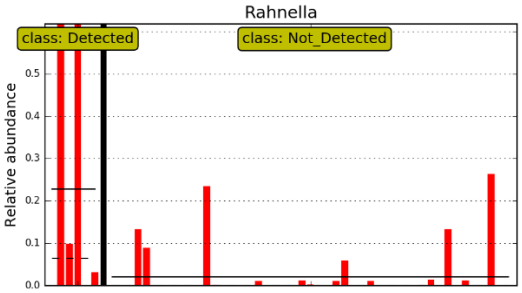

Lactococcus
class: Detected
class: Not_Detected
Relative abundance

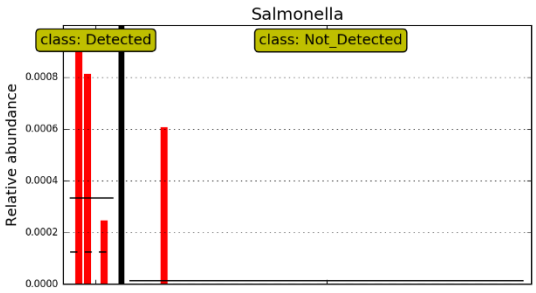

Leuconostoc
class: Detected
class: Not_Detected
Relative abundance

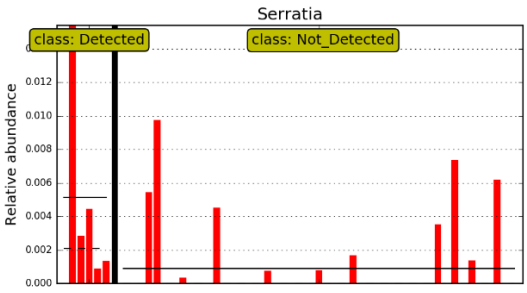

Morganella
class: Detected
class: Not_Detected
Relative abundance

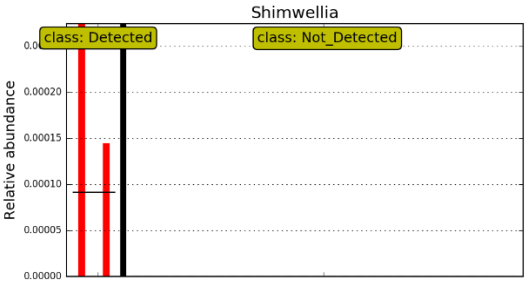

Pantoea
class: Detected
class: Not_Detected
Relative abundance

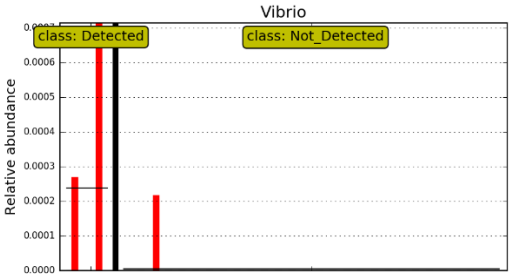

Pectobacterium
class: Detected
class: Not_Detected
Relative abundance

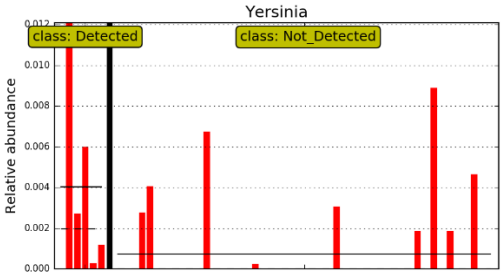

Rahnella
class: Detected
class: Not_Detected
Relative abundance


Salmonella
class: Detected
class: Not_Detected
Relative abundance


Serratia
class: Detected
class: Not_Detected
Relative abundance


Shimwellia
class: Detected
class: Not_Detected
Relative abundance
0.00020
0.00015
0.00010
0.00005
0.00000


Vibrio
class: Detected
class: Not_Detected
Relative abundance
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000


Yersinia
class: Detected
class: Not_Detected
Relative abundance
0.010
0.008
0.006
0.004
0.002
0.000

Appendix D. List of Isolates

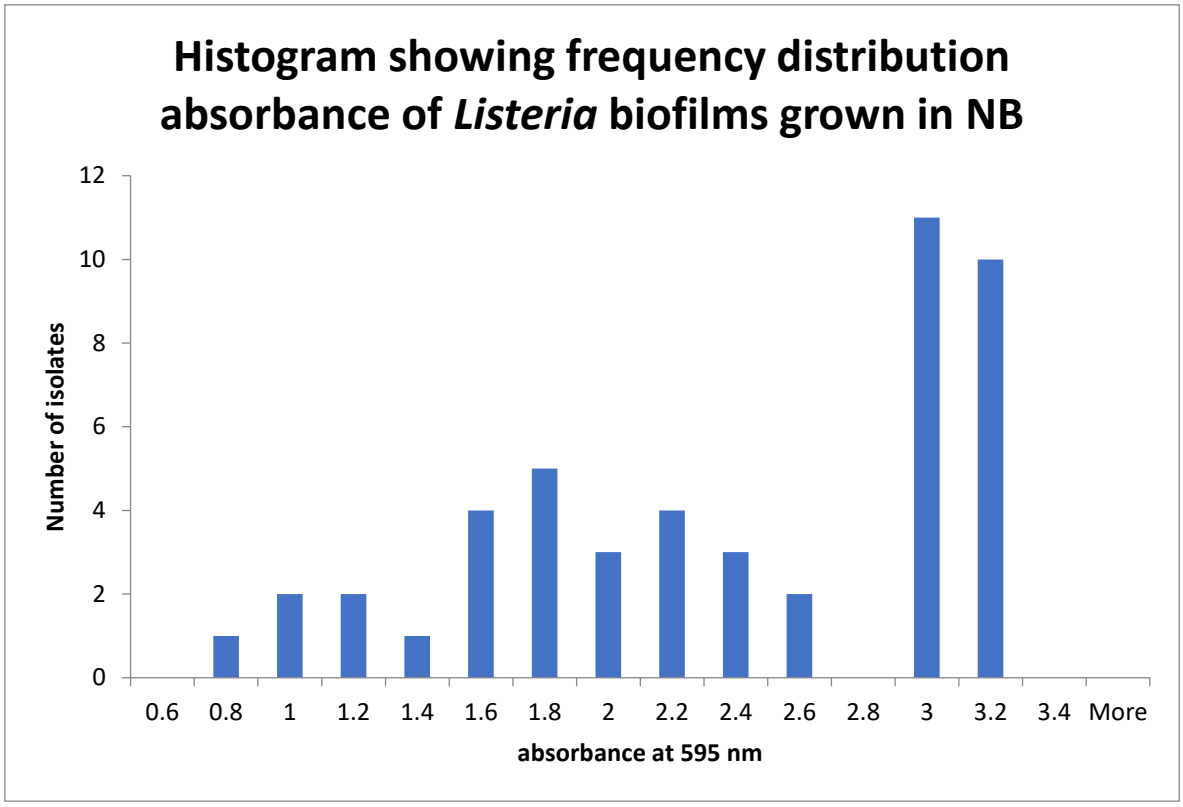
Name	Source	Data from	External Reference	Additional Metadata
clinical_SRR6518418	Clinical	NCBI SRA	SRR6518418	
clinical_SRR6518419	Clinical	NCBI SRA	SRR6518419	
clinical_SRR6798518	Clinical	NCBI SRA	SRR6798518	
clinical_SRR6798529	Clinical	NCBI SRA	SRR6798529	
clinical_SRR6798535	Clinical	NCBI SRA	SRR6798535	
clinical_SRR6798538	Clinical	NCBI SRA	SRR6798538	
clinical_SRR6798541	Clinical	NCBI SRA	SRR6798541	
clinical_SRR6798542	Clinical	NCBI SRA	SRR6798542	
clinical_SRR6798546	Clinical	NCBI SRA	SRR6798546	
clinical_SRR6798551	Clinical	NCBI SRA	SRR6798551	
clinical_SRR6798552	Clinical	NCBI SRA	SRR6798552	
clinical_SRR6798554	Clinical	NCBI SRA	SRR6798554	
clinical_SRR6798556	Clinical	NCBI SRA	SRR6798556	
clinical_SRR6798559	Clinical	NCBI SRA	SRR6798559	
clinical_SRR6798560	Clinical	NCBI SRA	SRR6798560	
clinical_SRR6798561	Clinical	NCBI SRA	SRR6798561	
clinical_SRR6798563	Clinical	NCBI SRA	SRR6798563	
clinical_SRR6798566	Clinical	NCBI SRA	SRR6798566	
clinical_SRR6798567	Clinical	NCBI SRA	SRR6798567	
clinical_SRR6798569	Clinical	NCBI SRA	SRR6798569	
clinical_SRR6798570	Clinical	NCBI SRA	SRR6798570	
clinical_SRR6798573	Clinical	NCBI SRA	SRR6798573	
clinical_SRR6798575	Clinical	NCBI SRA	SRR6798575	
clinical_SRR6798577	Clinical	NCBI SRA	SRR6798577	
clinical_SRR6798580	Clinical	NCBI SRA	SRR6798580	
clinical_SRR6798581	Clinical	NCBI SRA	SRR6798581	
clinical_SRR6798586	Clinical	NCBI SRA	SRR6798586	
clinical_SRR6798588	Clinical	NCBI SRA	SRR6798588	
clinical_SRR6798591	Clinical	NCBI SRA	SRR6798591	
clinical_SRR6798596	Clinical	NCBI SRA	SRR6798596	
clinical_SRR6798603	Clinical	NCBI SRA	SRR6798603	
clinical_SRR6798605	Clinical	NCBI SRA	SRR6798605	
clinical_SRR6798606	Clinical	NCBI SRA	SRR6798606	
clinical_SRR6798616	Clinical	NCBI SRA	SRR6798616	
clinical_SRR6798617	Clinical	NCBI SRA	SRR6798617	
clinical_SRR6798621	Clinical	NCBI SRA	SRR6798621	
clinical_SRR6798622	Clinical	NCBI SRA	SRR6798622	
clinical_SRR6798623	Clinical	NCBI SRA	SRR6798623	
clinical_SRR6798624	Clinical	NCBI SRA	SRR6798624	
clinical_SRR6798626	Clinical	NCBI SRA	SRR6798626	
clinical_SRR6798627	Clinical	NCBI SRA	SRR6798627	
clinical_SRR6798630	Clinical	NCBI SRA	SRR6798630	

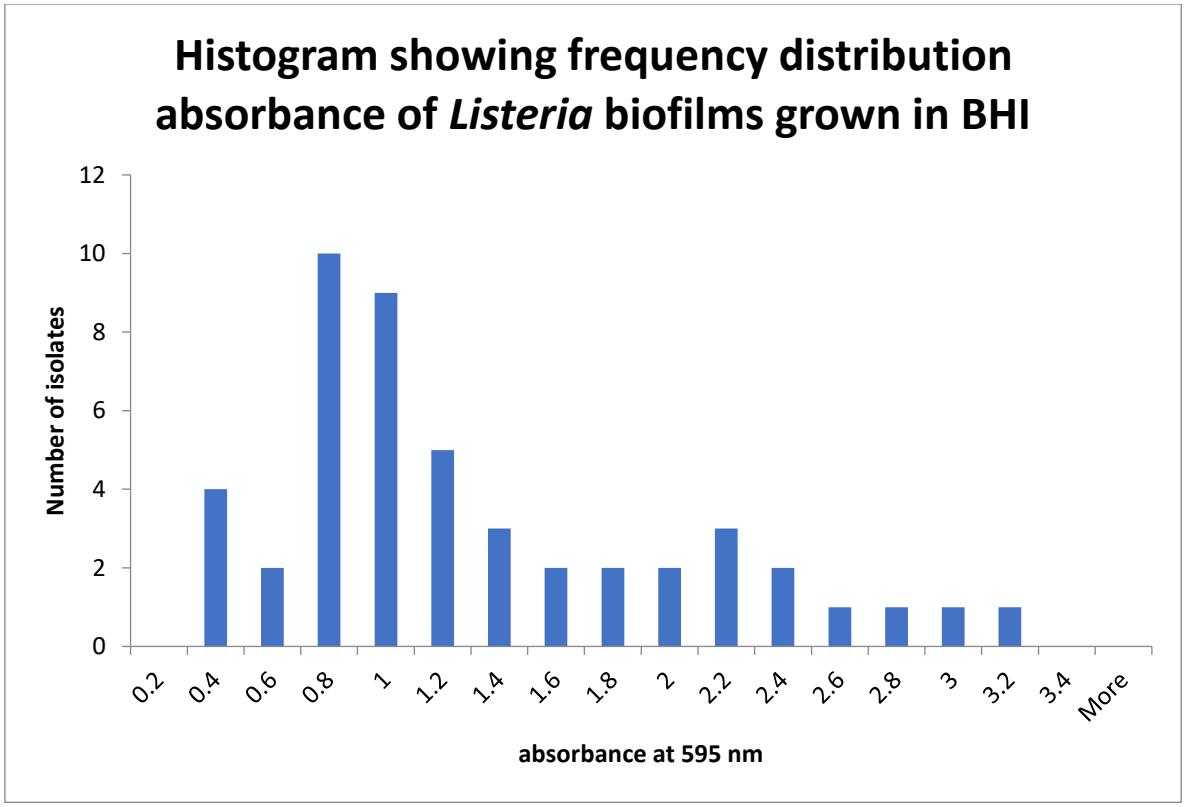
clinical_SRR6805082	Clinical	NCBI SRA	SRR6805082	
clinical_SRR6805087	Clinical	NCBI SRA	SRR6805087	
clinical_SRR6805285	Clinical	NCBI SRA	SRR6805285	
clinical_SRR6882060	Clinical	NCBI SRA	SRR6882060	
meat_SRR5182476	Meat	NCBI SRA	SRR5182476	
meat_SRR5344715	Meat	NCBI SRA	SRR5344715	
meat_SRS1908657	Meat	NCBI SRA	SRS1908657	
meat_SRS1908664	Meat	NCBI SRA	SRS1908664	
meat_SRS1908665	Meat	NCBI SRA	SRS1908665	
meat_SRS1908667	Meat	NCBI SRA	SRS1908667	
meat_SRS1908669	Meat	NCBI SRA	SRS1908669	
meat_SRS1908704	Meat	NCBI SRA	SRS1908704	
meat_SRS1908708	Meat	NCBI SRA	SRS1908708	
meat_SRS1908995	Meat	NCBI SRA	SRS1908995	
meat_SRS1908997	Meat	NCBI SRA	SRS1908997	
meat_SRS1909296	Meat	NCBI SRA	SRS1909296	
meat_SRS1909335	Meat	NCBI SRA	SRS1909335	
meat_SRS1909337	Meat	NCBI SRA	SRS1909337	
meat_SRS1909344	Meat	NCBI SRA	SRS1909344	
meat_SRS1909360	Meat	NCBI SRA	SRS1909360	
meat_SRS1909367	Meat	NCBI SRA	SRS1909367	
meat_SRS1909374	Meat	NCBI SRA	SRS1909374	
meat_SRS1909383	Meat	NCBI SRA	SRS1909383	
meat_SRS1909385	Meat	NCBI SRA	SRS1909385	
meat_SRS1909586	Meat	NCBI SRA	SRS1909586	
meat_SRS1909591	Meat	NCBI SRA	SRS1909591	
meat_SRS1909654	Meat	NCBI SRA	SRS1909654	
meat_SRS1909656	Meat	NCBI SRA	SRS1909656	
meat_SRS1909658	Meat	NCBI SRA	SRS1909658	
meat_SRS2048872	Meat	NCBI SRA	SRS2048872	
meat_SRS2048879	Meat	NCBI SRA	SRS2048879	
meat_SRS2048880	Meat	NCBI SRA	SRS2048880	
meat_SRS2048887	Meat	NCBI SRA	SRS2048887	
meat_SRS2048896	Meat	NCBI SRA	SRS2048896	
meat_SRS717409	Meat	NCBI SRA	SRS717409	
meat_SRS717411	Meat	NCBI SRA	SRS717411	
meat_SRS717414	Meat	NCBI SRA	SRS717414	
meat_SRS717415	Meat	NCBI SRA	SRS717415	
Reference_NCTC11994	Culture Collection	In house sequencing	-	
Reference_NCTC5214	Culture Collection	In house sequencing	-	
veg1	Vegetables	Smith et al. (2019)	Nlmo2	Spinach
veg2	Vegetables	Smith et al. (2019)	Nlmo3	Kale

veg3	Vegetables	Smith et al. (2019)	Nlmo4	Swab
veg4	Vegetables	Smith et al. (2019)	Nlmo5	Baby spinach
veg5	Vegetables	Smith et al. (2019)	Nlmo6	Red Leaf
veg6	Vegetables	Smith et al. (2019)	Nlmo7	Spinach
veg7	Vegetables	Smith et al. (2019)	Nlmo8	Baby spinach
veg8	Vegetables	Smith et al. (2019)	Nlmo9	Baby spinach
veg9	Vegetables	Smith et al. (2019)	Nlmo10	Spinach
veg10	Vegetables	Smith et al. (2019)	Nlmo13	Spinach
veg11	Vegetables	Smith et al. (2019)	Nlmo14	Beetroot
veg12	Vegetables	Smith et al. (2019)	Nlmo15	Peashoots
veg13	Vegetables	Smith et al. (2019)	Nlmo16	Spinach
veg14	Vegetables	Smith et al. (2019)	Nlmo18	Baby salad kale
veg15	Vegetables	In house sequencing	-	Baby Spinach
veg16	Vegetables	In house sequencing	-	Rocket
veg17	Vegetables	In house sequencing	-	Lollo Rosso
veg18	Vegetables	In house sequencing	-	Lollo Rosso
veg19	Vegetables	In house sequencing	-	Spinach
veg20	Vegetables	In house sequencing	-	Rocket
veg21	Vegetables	In house sequencing	-	Spring Greens
veg22	Vegetables	In house sequencing	-	Chinese Stir Fry
veg23	Vegetables	In house sequencing	-	Red onion
veg24	Vegetables	In house sequencing	-	Wild Rocket
veg25	Vegetables	In house sequencing	-	Spring Greens (sliced)
veg26	Vegetables	In house sequencing	-	Onions (diced)

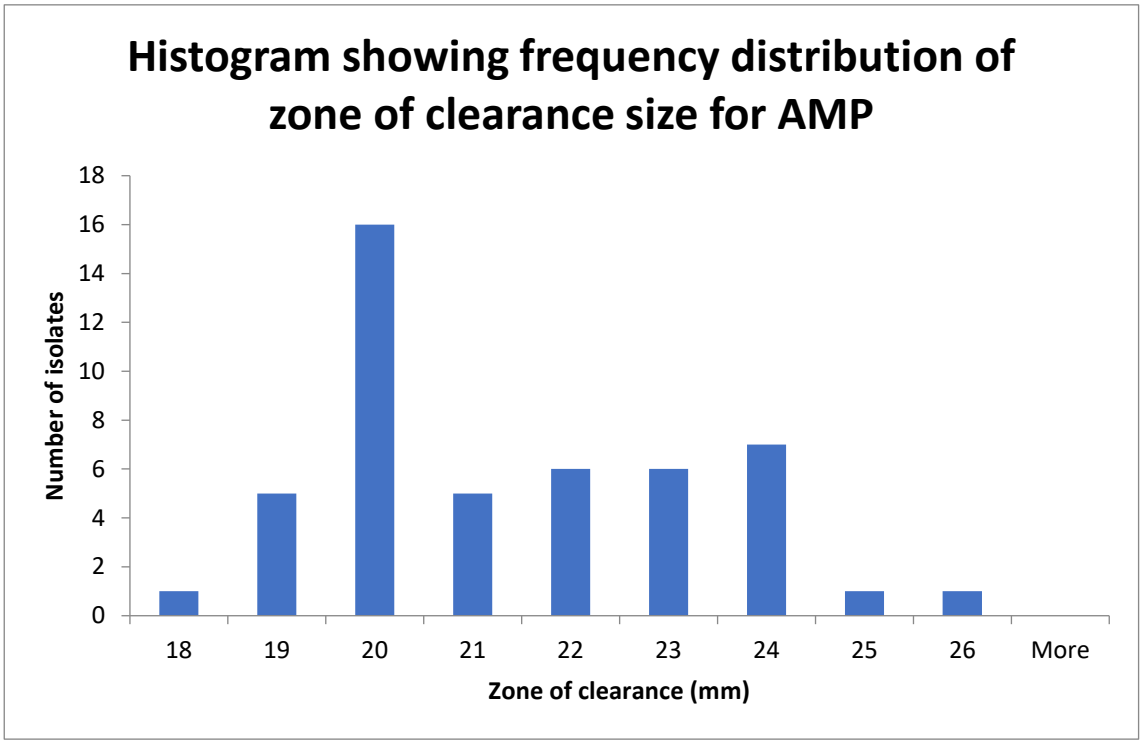
veg27	Vegetables	In house sequencing	-	Red Onion
veg28	Vegetables	In house sequencing	-	Spinach
veg29	Vegetables	In house sequencing	-	Spinach (unwashed)
veg30	Vegetables	In house sequencing	-	Mixed Salad
veg31	Vegetables	In house sequencing	-	Wild Rocket
veg32	Vegetables	In house sequencing	-	Spinach
veg33	Vegetables	In house sequencing	-	Mixed Salad
veg34	Vegetables	In house sequencing	-	Mixed Salad
veg35	Vegetables	In house sequencing	-	Wild Rocket
veg36	Vegetables	In house sequencing	-	Baby Spinach
veg37	Vegetables	In house sequencing	-	Mixed Salad
veg38	Vegetables	In house sequencing	-	Mixed Salad
veg39	Vegetables	In house sequencing	-	Spinach
veg40	Vegetables	In house sequencing	-	Red Chard
veg41	Vegetables	In house sequencing	-	Spinach
veg42	Vegetables	In house sequencing	-	Baby Spinach
veg43	Vegetables	In house sequencing	-	Spinach
veg44	Vegetables	In house sequencing	-	Spinach
veg45	Vegetables	In house sequencing	-	Spinach
veg46	Vegetables	In house sequencing	-	Spinach
veg47	Vegetables	In house sequencing	-	Baby Spinach
veg48	Vegetables	In house sequencing	-	Swab

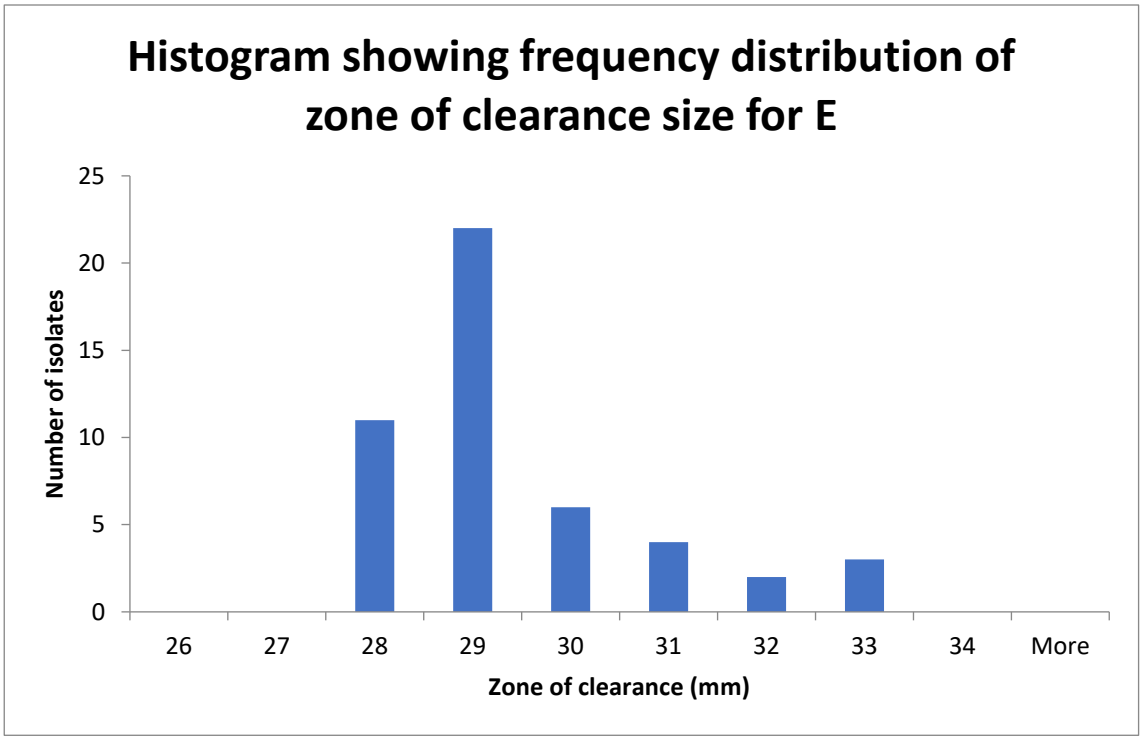
Appendix E. Histograms of biofilm formation

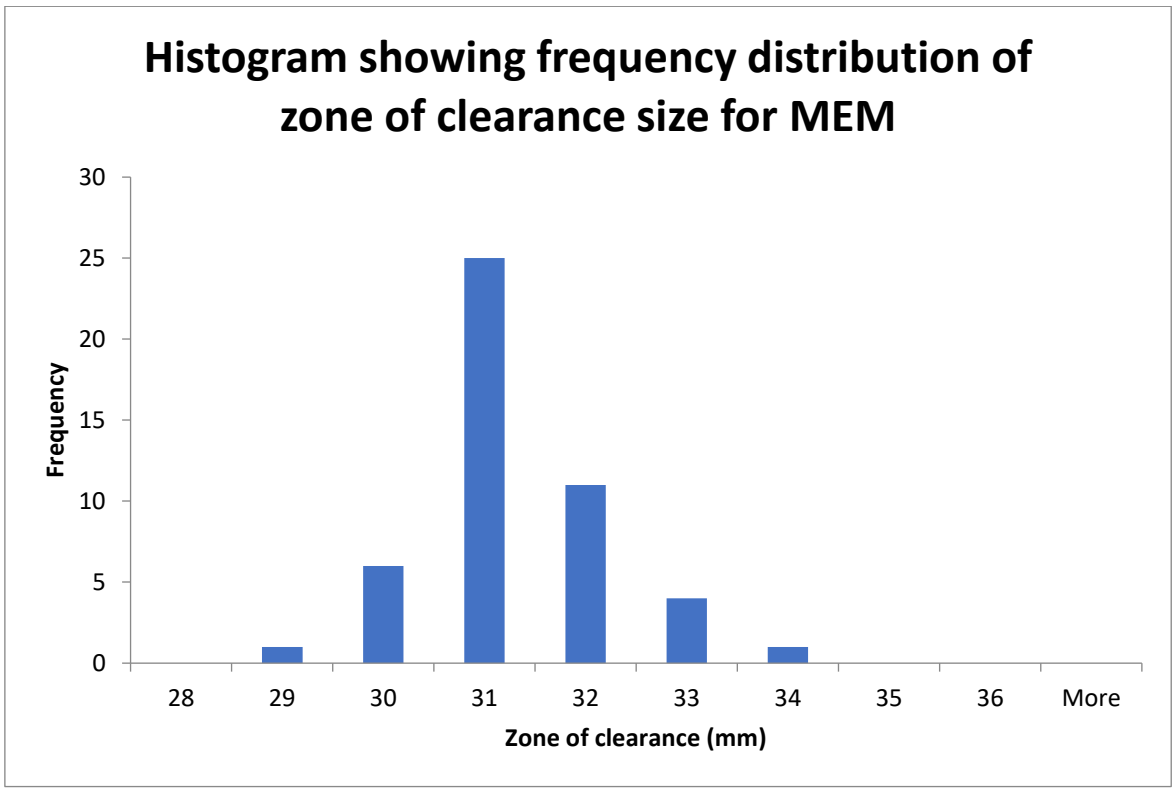


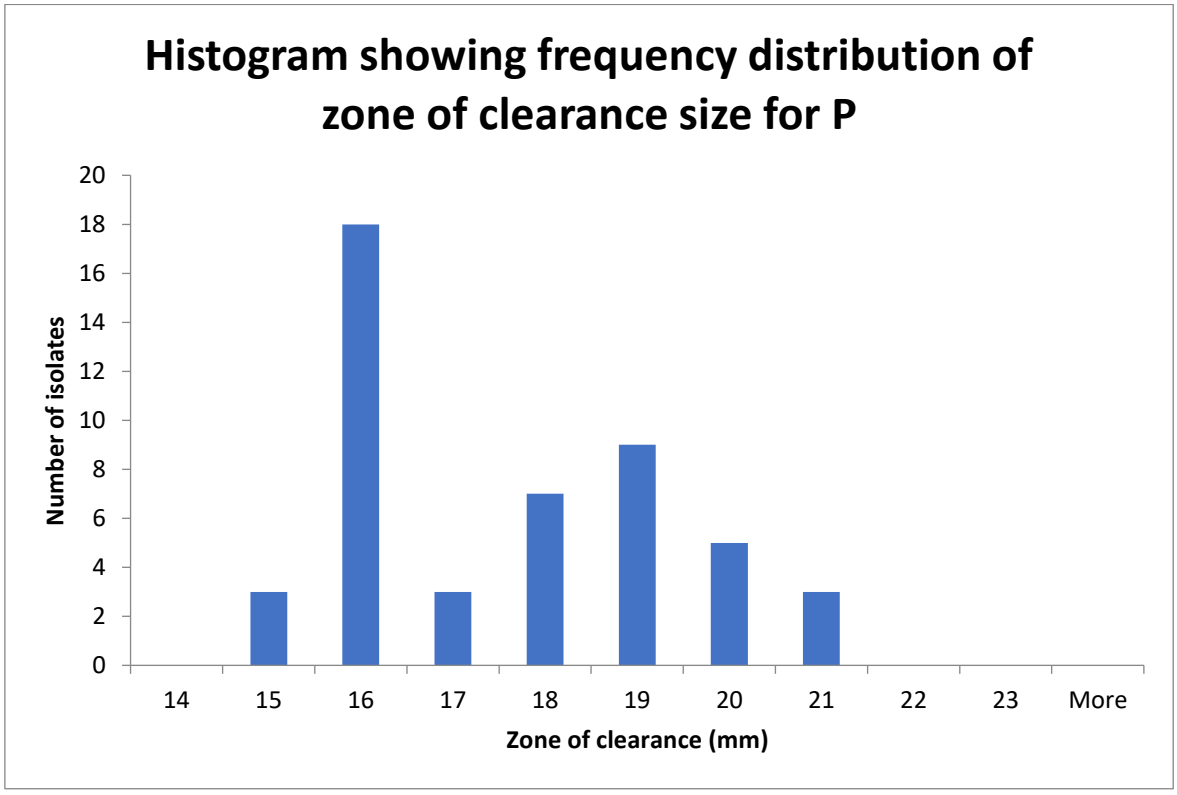


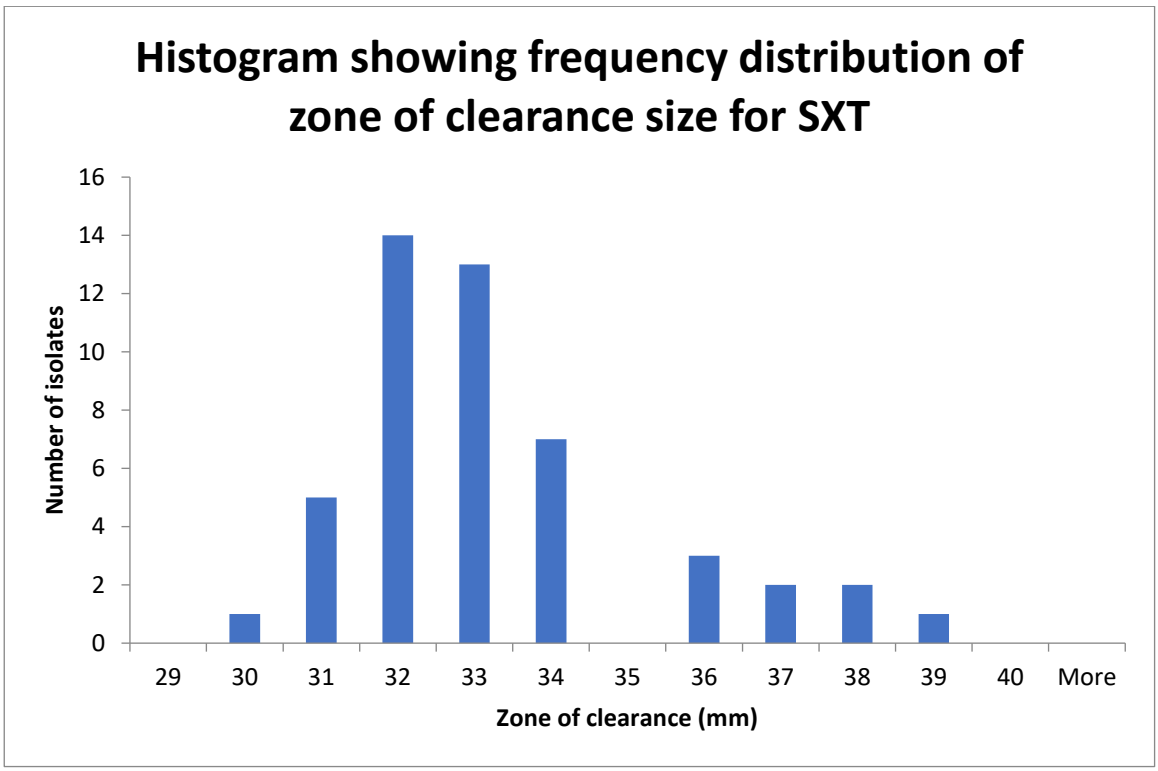
Appendix F. Histograms of antibiotic zone of clearance






Histogram showing frequency distribution of zone of clearance size for MEM
Frequency
30
25
20
15
10
5
0
28
29
30
31
32
33
34
35
36
More
Zone of clearance (mm)


Histogram showing frequency distribution of zone of clearance size for P
Number of isolates
20
18
16
14
12
10
8
6
4
2
0
14
15
16
17
18
19
20
21
22
23
More
Zone of clearance (mm)


Histogram showing frequency distribution of zone of clearance size for SXT
Number of isolates
16
14
12
10
8
6
4
2
0
29
30
31
32
33
34
35
36
37
38
39
40
More
Zone of clearance (mm)

Appendix G. Pyseer Genes of Interest

i) Genes associated with isolation from vegetables

gene	hits	maxp	avg_af	avg_maf	avg_beta
int	1143	17.30539	0.17126	0.17126	0.828628
cds2325	546	15.29073	0.173886	0.173886	0.824284
cds2326	415	14.11748	0.174308	0.174308	0.822265
cds2310	365	15.29073	0.16891	0.16891	0.817989
cds2329	300	14.11748	0.183077	0.183077	0.790897
cds1801	294	15.27409	0.175671	0.175671	0.812306
cds2323	255	14.37161	0.165813	0.165813	0.828929
cds2328	230	14.11748	0.202709	0.202709	0.71037
cds2273	229	13.57675	0.177978	0.177978	0.826406
cds2324	204	13.57675	0.172142	0.172142	0.818074
cds1350	158	14.11748	0.173285	0.173285	0.812854
cds1223	157	14.11748	0.173573	0.173573	0.818866
cds1723	155	14.11748	0.18351	0.18351	0.809497
ftsA	153	14.11748	0.174734	0.174734	0.809007
cds1739	151	14.11748	0.171689	0.171689	0.821053
cds499	151	14.11748	0.169841	0.169841	0.836219
cds639	149	14.11748	0.174619	0.174619	0.82053
cds1580	143	14.11748	0.183471	0.183471	0.824916
cds2580	141	14.11748	0.16508	0.16508	0.818383
cds129	139	13.57675	0.168317	0.168317	0.827158
cds2337	136	13.80688	0.173743	0.173743	0.819272
cds2274	135	14.11748	0.167691	0.167691	0.837304
cds2394	133	16.25337	0.176015	0.176015	0.822105
ruvA	132	14.11748	0.174614	0.174614	0.824364
serC	120	14.11748	0.181683	0.181683	0.817208
cds1736	114	14.31876	0.179193	0.179193	0.820991
cds60	108	13.57675	0.20137	0.20137	0.800417
cds585	105	14.11748	0.173914	0.173914	0.807333
cds517	95	14.23136	0.164579	0.164579	0.830042
cds2693	91	13.57675	0.190659	0.190659	0.817066
cds2331	84	14.93554	0.163146	0.163146	0.82125
cds2317	76	13.57675	0.175618	0.175618	0.810618
cds2400	75	14.11748	0.203017	0.203017	0.813773
qoxC	73	14.11748	0.182233	0.182233	0.832014
cds2318	70	13.69465	0.162486	0.162486	0.853643
cds887	70	14.68194	0.180886	0.180886	0.820843
cds2315	64	13.57675	0.168203	0.168203	0.823797
cds2312	62	13.80688	0.184516	0.184516	0.819323

cds460	61	13.57675	0.181349	0.181349	0.798541
cds1136	58	14.16178	0.16781	0.16781	0.821983
cds2327	56	14.11748	0.175268	0.175268	0.816357
cds2311	52	13.9431	0.170987	0.170987	0.828558
cds2330	48	13.57675	0.174167	0.174167	0.805437
cds496	44	13.57675	0.16	0.16	0.824432
comK'	44	15.29073	0.180432	0.180432	0.800955
cds2399	41	13.69465	0.171927	0.171927	0.843537
ssb	34	13.27984	0.219	0.219	0.777882
cds159	31	13.57675	0.198484	0.198484	0.843548
cds2305	14	13.27984	0.194429	0.194429	0.781643
cds2398	13	13.27984	0.162231	0.162231	0.795538
cds456	7	13.27984	0.173714	0.173714	0.768286
cds431	7	12.18111	0.370286	0.370286	0.622
cds2334	5	12.87615	0.2986	0.2986	0.6066
inlA	4	13.27984	0.1655	0.1655	0.75825
cds919	4	13.9431	0.42875	0.42875	0.71475
thrS	4	11.03058	0.45375	0.45375	0.6985
cds2015	4	9.896196	0.5	0.5	0.62825
cds1979	4	10.7122	0.4945	0.4945	0.67425
cds539	3	10.35655	0.453667	0.453667	0.641
cds2592	3	10.54363	0.351	0.351	0.699333
cds106	3	13.41341	0.474333	0.474333	0.74
cds2302	3	8.756962	0.274333	0.274333	0.526667
cds1002	3	9.543634	0.407667	0.407667	0.678
cds2581	3	11.39147	0.464333	0.464333	0.792667
cds1207	3	8.178486	0.479333	0.479333	0.566
cds130	2	13.48149	0.4575	0.4575	0.675
cds1719	2	8.779892	0.4305	0.4305	0.583
cds556	2	11.49757	0.462	0.462	0.758
cds2	2	11.48017	0.3305	0.3305	0.7345
cds1096	2	8.614394	0.281	0.281	0.621
cds147	2	9.415669	0.45	0.45	0.543
cds2742	2	10.86646	0.4385	0.4385	0.7155
cds2450	2	8.739929	0.35	0.35	0.5855
cds433	2	8.059484	0.5115	0.4885	0.625
cds65	2	8.463442	0.4	0.4	0.584
purM	2	9.102923	0.4615	0.4615	0.5915
cds723	2	11.68613	0.481	0.481	0.8275
cds398	2	10.53611	0.477	0.477	0.742
glyQ	2	9.032452	0.3965	0.3965	0.5505
cds1679	2	14.04001	0.4305	0.4305	0.773
purD	2	10.24109	0.4805	0.4805	0.7255
bvrA	2	12.54821	0.4845	0.4845	0.759
cds941	2	8.879426	0.427	0.427	0.6005
cds420	2	10.95078	0.4	0.4	0.7005

cds319	2	8.806875	0.4805	0.4805	0.628
cds2180	2	10.05306	0.442	0.442	0.681
cds2003	2	8.220404	0.627	0.373	0.648
cds1763	2	11.86967	0.515	0.485	0.657
cds1355	2	8.271646	0.442	0.442	0.6985
cds2016	2	12.18842	0.4195	0.4195	0.6865
cds2279	2	7.866461	0.1885	0.1885	0.629
cds63	2	8.688246	0.5575	0.4425	0.5515
cds2320	2	8.982967	0.1885	0.1885	0.66
atpI	2	13.56384	0.4695	0.4695	0.825
cds1821	2	14.65956	0.454	0.454	0.844
codY	2	12.75449	0.477	0.477	0.7135
fliD	2	12.81248	0.423	0.423	0.764
cds824	2	8.853872	0.477	0.477	0.6195
cds2674	2	10.47366	0.492	0.492	0.708
cds2783	2	11.31158	0.327	0.327	0.6725
cds327	2	9.477556	0.458	0.458	0.6625
ilvD	2	8.946922	0.296	0.296	0.656
cds2027	2	10.04769	0.4575	0.4575	0.663
cds720	2	8.876148	0.327	0.327	0.6655
cds2174	2	9.527244	0.423	0.423	0.688
cds835	2	11.38091	0.45	0.45	0.778
cds457	1	11.61798	0.154	0.154	0.842
cds1102	1	9.247952	0.169	0.169	0.709
cds2256	1	9.243364	0.315	0.315	0.65
cds1119	1	7.761954	0.277	0.277	0.657
gltX	1	9.02641	0.385	0.385	0.656
cds2308	1	8.723538	0.215	0.215	0.628
cds218	1	14.27654	0.477	0.477	0.874
cds1583	1	7.838632	0.292	0.292	0.588
purN	1	8.320572	0.385	0.385	0.656
cds2289	1	9.675718	0.3	0.3	0.549
cds2680	1	7.735182	0.423	0.423	0.558
cds2599	1	8.551294	0.415	0.415	0.567
cds266	1	9.224026	0.4	0.4	0.663
cds2376	1	9.104025	0.277	0.277	0.733
cds1093	1	9.411168	0.431	0.431	0.605
cds802	1	8.25649	0.562	0.438	0.473
cds1289	1	8.801343	0.346	0.346	0.688
cds2127	1	8.958607	0.408	0.408	0.669
cds1422	1	10.54516	0.446	0.446	0.73
cds821	1	11.16178	0.408	0.408	0.726
cds2297	1	8.054039	0.231	0.231	0.591
cds479	1	9.492144	0.423	0.423	0.675
purB	1	9.192465	0.454	0.454	0.682
cds2004	1	8.801343	0.385	0.385	0.636

cds1647	1	10.23062	0.292	0.292	0.725
cds839	1	9.793174	0.515	0.485	0.622
cds279	1	13.61979	0.169	0.169	0.895
murD	1	8.5867	0.331	0.331	0.648
cds315	1	8.492144	0.308	0.308	0.629
cds1960	1	9.396856	0.308	0.308	0.713
cds2785	1	8.486782	0.423	0.423	0.525
cds316	1	8.879426	0.354	0.354	0.648
cds785	1	8.74958	0.323	0.323	0.687
cbiP	1	10.1209	0.308	0.308	0.703
cds2346	1	7.742321	0.408	0.408	0.517
lysS	1	8.112946	0.485	0.485	0.638
cds250	1	8.111259	0.315	0.315	0.635
cds1500	1	9.291579	0.508	0.492	0.66
tkt	1	8.920819	0.285	0.285	0.723
cds464	1	11.29757	0.285	0.285	0.785
cds2678	1	10.50446	0.269	0.269	0.781
cds506	1	8.928118	0.469	0.469	0.685
cds363	1	7.815309	0.392	0.392	0.517
cds1211	1	7.787812	0.569	0.431	0.595
cds2303	1	8.701147	0.208	0.208	0.611
cds253	1	8.195861	0.377	0.377	0.61
cds2321	1	9.156767	0.169	0.169	0.71
cds2681	1	13.24489	0.385	0.385	0.754
phoR	1	9.394695	0.346	0.346	0.652
cds786	1	9.812479	0.577	0.423	0.595
cds382	1	9	0.477	0.477	0.66
cds761	1	10.16877	0.469	0.469	0.708
cds66	1	7.818156	0.315	0.315	0.623
cds2030	1	11.02733	0.392	0.392	0.708
cds2818	1	9.432974	0.385	0.385	0.676
cds1204	1	11.67162	0.469	0.469	0.812
gbuA	1	11.49485	0.477	0.477	0.805
cds2500	1	8.653647	0.292	0.292	0.684
cds1099	1	13.24565	0.185	0.185	0.799
cds469	1	10.15058	0.315	0.315	0.662
cds1235	1	10.27165	0.469	0.469	0.699
cds518	1	9.427128	0.269	0.269	0.713
cds2640	1	11.70774	0.485	0.485	0.798
cds2755	1	8.492144	0.369	0.369	0.657
trmE	1	8.492144	0.369	0.369	0.657
cds289	1	9.189767	0.692	0.308	0.683
cds2277	1	10.21254	0.169	0.169	0.742
cds1923	1	9.241845	0.415	0.415	0.644
cds1114	1	11.07314	0.162	0.162	0.787
cds131	1	8.191789	0.354	0.354	0.564

cds2858	1	9.9914	0.438	0.438	0.714
cds742	1	8.114639	0.308	0.308	0.63
cds1035	1	9.767004	0.285	0.285	0.697
cds1030	1	10.0804	0.562	0.438	0.609
cds1307	1	7.935542	0.508	0.492	0.62
dapF	1	7.761954	0.431	0.431	0.626
pheS	1	7.954677	0.331	0.331	0.613
cds2840	1	9.002614	0.408	0.408	0.673
cds2309	1	10.21538	0.162	0.162	0.742
hisA	1	7.978811	0.638	0.362	0.642
cds1201	1	12.55596	0.392	0.392	0.733
cds74	1	9.14813	0.531	0.469	0.61
cds2434	1	11.80967	0.469	0.469	0.784
cds276	1	11.51856	0.254	0.254	0.736
cadA	1	8.28567	0.385	0.385	0.614
cds2654	1	8.679854	0.323	0.323	0.599
cds1376	1	12.37675	0.323	0.323	0.78
gyrB	1	8.630784	0.346	0.346	0.651
cds898	1	9.835647	0.577	0.423	0.587
menB	1	8.112946	0.631	0.369	0.537
cds78	1	8.283162	0.446	0.446	0.585
cds2584	1	9.806875	0.408	0.408	0.617
cds1117	1	7.978811	0.308	0.308	0.588
pyrG	1	13.41341	0.469	0.469	0.88
cds2849	1	10.33536	0.454	0.454	0.73
acpP	1	8.399027	0.423	0.423	0.655
cds2052	1	8.308035	0.362	0.362	0.581
cds2448	1	7.790485	0.5	0.5	0.656
cds849	1	11.76447	0.431	0.431	0.758
cds586	1	14.41454	0.454	0.454	0.861
cds516	1	10.39686	0.585	0.415	0.616
cds2278	1	8.003051	0.231	0.231	0.574
cds2050	1	8.420216	0.285	0.285	0.676
cds486	1	9.425969	0.338	0.338	0.688
cds2594	1	8.850781	0.423	0.423	0.633
cds1865	1	8.266001	0.277	0.277	0.673
cds323	1	9.546682	0.492	0.492	0.613
cds1115	1	10.55129	0.138	0.138	0.792
cds1671	1	8.910095	0.308	0.308	0.642
cds2679	1	11.25885	0.492	0.492	0.755
glyS	1	7.739929	0.408	0.408	0.55
cds1577	1	8.450997	0.408	0.408	0.62

ii) Genes associated with isolation from clinical samples

gene	hits	maxp	avg_af	avg_maf	avg_beta
cds460	44	12.69037	0.684659	0.315341	0.517909
cds456	41	10.05355	0.708073	0.291927	0.517902
cds2400	8	13.94692	0.656	0.344	0.58775
gsaB	2	7.939302	0.738	0.262	0.517
cds479	2	10.05355	0.731	0.269	0.571
cds459	2	7.982967	0.654	0.346	0.478
cds2640	2	9.806875	0.869	0.131	0.743
prfA	1	8.165579	0.831	0.169	0.614
cds2592	1	8.163676	0.638	0.362	0.479
cds158	1	8.616185	0.792	0.208	0.582
cds1226	1	7.995679	0.869	0.131	0.676