THE UNIVERSITY
of EDINBURGH

EHEC O157 from A to T: EHEC O157:H7 epidemiology supplemented with long-read sequencing.

Sharif Shaaban Muhammad Shaaban

PhD Infection and Immunity – The University of Edinburgh – 2019

Table of Contents

Tables and Figures

Appendix

Key Abbreviations

AMR	Antimicrobial Resistance
BLAST	Basic Local Alignment Search Tool
bp	Base Pair
BWA	Burrows Wheeler Alignment
CDC	Center for Disease Control
DNA	Deoxyribonucleic Acid
ECDC	European Center for Disease Control
EHEC	Enterohaemorrhagic *Escherichia coli*
GB	GigaByte
Gbp	Giga-base pair
GI	Gastrointestinal
HPS	Health Protection Scotland
HUS	Haemolytic Uremic Syndrome
IS	Insertion Sequence
Kbp	Kilo-base pair
LCR	Large Chromosomal Rearrangement
LEE	Locus of Enterocyte Effacement
LPS	Lipopolysaccharide
MLST	Multi-Locus Sequence Typing
MLVA	Multi-Locus Variable-number tandem repeat Analysis
MOST	Metric Oriented Sequence Typer
NGS	Next-Generation Sequencing
O157	EHEC O157:H7
OI	O Island
OS	Operating System

PacBio	Pacific Biosciences
PAI	Pathogenicity Island
PCR	Polymerase Chain Reaction
PFGE	Pulse Field Gel Electrophoresis
PH	Public Health
PHA	Public Health Agency
PHE	Public Health England
PHW	Public Health Wales
PT	Phage Type
SBS	Sequencing By Synthesis
SERL	Scottish *Escherichia coli* Reference Laboratory
SMRT	Single-Molecule Real-Time
SNP	Single Nucleotide Polymorphism
ST	Sequence Type
STEC	Shiga Toxin producing *Escherichia coli*
Stx	Shiga Toxin
USDA	United States Department of Agriculture
VTEC	Verotoxin producing *Escherichia coli*
WGS	Whole Genome Sequencing
ZMW	Zero-Mode Waveguide

Declaration of Own Work

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Parts of this work have been published in:

S. Shaaban, L. A. Cowley, S. P. McAteer, C. Jenkins, T. J. Dallman, J. L. Bono, D.L. Gally. Evolution of a zoonotic pathogen: investigating prophage diversity in enterohaemorrhagic Escherichia coli O157 by long-read sequencing. *Mgen*, 2016

Date and Signature:

Sharif S. M. Shaaban

Glasgow, UK, 22/09/2019

Main body final word count (with in text citations but excluding bibliography):

32353 Words

Acknowledgements

“Life is a journey and not a destination.” - Lynn H Hough

Here I am standing at the end of this journey that has been my PhD degree. An expedition that was started nearly five years ago with its fair share of twists, turns, challenges, and deep pits, and yet as I am writing the final words of this work, all I feel is sadness that the journey is over rather than joy at (potentially) having reached this destination.

While there are too many people I need to thank for allowing me to reach this point, some need to be singled out: my parents (to whom I sincerely apologize for taking too long), Jehanne Tewfik and Muhammad Shaaban who with their constant (loving) nagging never allowed me to give up on this adventure, my supervisors, David Gally and Tim Dallman who despite my relentless questions and digressions never gave up on me, my collaborators throughout this project, Lesley Allison and Anne Holmes without whom I would have been unable, and unaware of the possibility, to take my next step in the field of public health, my cousin, Yousef Hesham who will never realise how much his support throughout everything meant, and too many of my friends to actually name them all who made my lows bearable and my highs even greater. To them and everyone else who was involved directly or indirectly in this PhD, I thank you from the bottom of my heart. I hope to see you all for my next journey.

Sharif S. M. Shaaban

Glasgow, UK, 22/09/2019

“I’m going on an adventure.” – J R R Tolkien; The Hobbit

Abstract

Enterohaemorrhagic *Escherichia coli* O157:H7 (EHEC O157) is a key zoonotic pathogen responsible for large food-borne outbreaks worldwide. Whole genome sequencing is a relatively novel technology being utilised by Public Health agencies to determine isolate relationship and inform outbreak investigations. However, the main implementation of whole genome sequencing currently utilises “short-read” sequencing which fails to obtain complete information on prophages and genome structure due to the presence of multiple repeat regions in EHEC O157 genomes.

In collaboration with Public Health England, this research helped deploy short-read sequencing approaches for routine use at the Scottish Escherichia coli Reference Laboratory (SERL). This has allowed the SERL and affiliated Scottish epidemiologists to better determine whether isolates are related and trace the source of outbreaks. The “long-read” Pacific Biosciences (PacBio) sequencing platform was then used to analyse the complete prophage content (bacteriophage DNA integrated in the chromosome) of a subset of strains. Specifically, the analysis took an in-depth look at prophages encoding the main Shiga toxins (Stx) responsible for the serious pathology associated with EHEC O157 infections. In addition, the sequencing method allowed the observation of large chromosomal rearrangements (LCRs), potentially mediated by areas of homologies present in the prophage population. The significance of such LCRs is still being investigated but the genome plasticity may act to allow the bacterial strain to ‘switch’ phenotypes for niche adaptation.

The potential of using “long-read” methods alongside routine “short-read” sequencing of EHEC O157 for public health benefit was investigated and its value demonstrated for outbreaks. For example, by enabling a more accurate prediction of the host/source attribution of an infection strain based on analysis of the Stx-encoding prophage within the isolate. While such

approaches show considerable promise, costs and accuracy issues (depending on the platform, PacBio vs Oxford Nanopore MinIon) will need to be surmounted, and the underlying biology studied further, before their use could usurp more high throughput “short-read” methods.

Lay Summary

Enterohaemorrhagic *Escherichia coli* O157:H7 (EHEC O157) is one of the main types of E. coli which can cause disease. Its main symptoms are bloody diarrhoea, but in some cases, it can lead to kidney failure, and even death. Most outbreaks are spread through food such as raw meat or unwashed vegetables that may have come in contact with animal's faecal matter. Therefore, it is an organism which is tracked by Public Health agencies and studied by academics to be able to quickly respond to an outbreak or infection.

One of the key aims of my project was to deploy a new typing technology at the Scottish *E. coli* Reference Laboratories (SERL), in order to improve their identification of outbreaks. This technology is called whole genome sequencing, which allows for the DNA signature of the infectious bacteria to be determined. This allows us to compare bacteria at a previously unattainable resolution, permitting us to rapidly identify and determine whether a received isolate is part of an outbreak and also helps in detecting the source of the outbreak more rapidly. This technology, which is becoming common across the world and public health institutions, was successfully deployed at SERL in collaboration with Public Health England (PHE).

However, this methodology has one main drawback in that it relies on a type of DNA sequencing which outputs only short DNA reads. Therefore, when putting them all back together, it is unable to decipher the complete DNA sequence of an EHEC O157 strain and will divide it in multiple parts. This is due to the DNA of EHEC O157 having repeated sequences often larger than the length of these DNA reads. While this does not necessarily affect the tracking of isolates, the information usually prevents us generating complete sequences for prophage regions (bits of viral DNA integrated into the bacterial DNA) that encode the main toxins responsible for the serious disease caused by EHEC O157. Fortunately, long-read sequencing technologies have been rapidly improving and are becoming commonly used

as costs decrease. These allow for much longer DNA reads to be sequenced. This technology was applied in this research to generate a number of completely assembled DNA sequences of EHEC O157 isolates, for which the whole DNA strand is sequenced from start to end with no gaps or breaks. This information has allowed me to study prophages and chromosomal structural variation that would not have been possible using the traditional short-read methods.

Using long-read sequencing data results were generated demonstrating that prophages could be used to improve outbreak tracking capabilities. Furthermore, long-read sequencing data was able to help further understand situations that were ambiguous when using short-read sequencing. However, most interestingly was the observation that the DNA of EHEC O157 appears to be prone to large rearrangements most likely mediated by repeated DNA code present in prophages. While it is unclear whether these rearrangements have an effect on the phenotype of the bacteria. i.e. its physical characteristics such as the severity of the infection, or the survival potential of the bacteria; preliminary data generated by further studies building up on this project appear to indicate so.

1 Introduction

This thesis was written with result chapters (**Chapters 2, 3,** and **4**) having their own introduction and discussion. Due to the nature of this project, and the importance of software, certain method sections will also include short introductions to tools. Therefore, to avoid repetition, this introduction will focus on information that is relevant to all the subsequent chapters, thus allowing for each result chapter to be its own standalone unit.

1.1 Enterohaemorrhagic Escherichia coli

Escherichia coli (or *E. coli*) is typically a commensal species of bacteria that is generally found in the gastrointestinal (GI) system of warm-blooded organisms (Kaper, Nataro, and Mobley 2004). It is a rod-shaped Gram-negative bacterium (Lim, Yoon, and Hovde 2010). However, there are subsets of the species that can be pathogenic in animals and/or humans (Kaper, Nataro, and Mobley 2004). The highly specific clonal groups are known as pathotypes. Their virulence is typically associated with different combinations of virulence factors acquired by horizontal gene transfer (Kaper, Nataro, and Mobley 2004; Dallman et al. 2015). The main pathotypes responsible for GI infections are:

- Enteropathogenic *E. coli* (EPEC)
- Enterohaemorrhagic *E. coli* (EHEC)
- Enterotoxigenic *E. coli* (ETEC)
- Enteroaggregative *E. coli* (EAEC)
- Enteroinvasive *E. coli* (EIEC)
- Diffusely Adherent *E. coli* (DAEC)

Other common pathotypes include Uropathogenic *E. coli* (UPEC), and other Extraintestinal Pathogenic *E. coli* (ExPEC) such as meningitis-associated *E. coli* (MNEC), and Avian Pathogenic *E. coli* (APEC) (Kaper, Nataro, and Mobley 2004).

This body of work focuses on EHEC also known as Shiga-toxin producing *E. coli* (STEC), or Vero-toxin producing *E. coli* (VTEC). EHEC is a zoonotic pathogen for which cattle are considered the main animal reservoir (Armstrong, Hollingsworth, and Morris 1996; Chase-Topping et al. 2008), although other ruminants can be implicated, such as sheep (Lim, Yoon, and Hovde 2010). EHEC, as with other *E. coli*, can colonise the GI tract of animal hosts and then be transmitted to humans directly through the faecal-oral route or indirectly through contaminated food (meats, dairy products or salad and vegetable produce) and drink, for example private water supplies (Chase-Topping et al. 2008). EHEC infection can result in diarrhoea or bloody diarrhoea. The most serious outcome is Haemolytic Uremic Syndrome (HUS) which can lead to kidney failure and sometimes can be fatal (Byrne et al. 2015). More serious pathology is usually observed in infants, immunocompromised individuals, and the elderly (Byrne et al. 2015).

Typical EHEC are generally defined on the basis that they encode a type 3 secretion system and Shiga toxins (Stx), as well as some association with human disease (Reid et al. 2000). However, in certain cases EHEC can exhibit features of other pathotypes, for example there was a large ‘atypical’ EHEC outbreak in Northern Germany in 2011 for which the associated isolates did not encode a type 3 secretion system and instead colonised gastrointestinal epithelium using adhesins which are typically expressed by another pathotype, enteroaggregative *E. coli* (EAEC) (Kaper, Nataro, and Mobley 2004; Karch et al. 2012). EHEC is further diversified by the fact that multiple serotypes (typing of the bacterium based on the antisera markers present on its membrane, introduced further in **Chapter 2**) of *E. coli* can be defined as EHEC. The key serotypes that are considered the main threat to human health are: O157, and O26. However, the serotype itself does not account for the pathogenicity of the organism. The combination of virulence determinants present in specific strains of the serotype and how they are regulated due to the habitat(s) of the strain will mediate its pathogenicity. If we consider serotypes to have evolved from common ancestors, it can therefore be assumed that when two serotypes diverge from their ancestor,

they may acquire virulence factors which become a common profile for this specific serotype if it is clonal (such as EHEC O157) (divergent evolution). Conversely, acquisition in different backgrounds can result in similar pathogens (convergent evolution).

In the UK, the EHEC serotype most commonly responsible for human infections is EHEC O157:H7 (meaning its O antigen is variant 157, and its H antigen is variant 7) (Chase-Topping et al. 2008; Dallman et al. 2015; Holmes et al. 2018). EHEC O157:H7 (referred to as EHEC O157 in this thesis) is responsible for outbreak events in countries worldwide and can often be found in news headlines regarding new large foodborne outbreaks (Chase-Topping et al. 2008). As mentioned above, other EHEC serotypes can lead to human infections, with non-EHEC O157 infections becoming more common in Scotland than EHEC O157 infections in recent years (Chase-Topping et al. 2008). In part this is down to increased surveillance and it remains the case that EHEC O157 is a critical serotype responsible for a high proportion of the more serious human infections, especially in the UK, North and South America.

1.2 Shiga Toxins

Stx are a toxin family which inhibit protein synthesis within target cells. They historically have been called verotoxins, due to the high susceptibility of Vero cells to these toxins (Pacheco and Sperandio 2012); Vero cells were derived from kidney epithelial cells. The toxicity of bacterial supernatants on Vero cells (Vero cell assays) have traditionally been used as a proxy to understand a strains potency, although there is now evidence that different Stx subtypes behave quite differently on these cells (Melton-Celsa 2014). Interestingly, these toxins also offer an explanation as to why cattle is an asymptomatic carrier of the bacteria. Stx targets globotriaosylceramide (Gb3) which is a surface component of non-epithelial endothelial cells in humans but not in cattle (Pruimboom-Brees et al. 2000; Kaper, Nataro, and Mobley 2004). The binding of the toxin to Gb3 allows the toxin to enter the cell and inhibit protein synthesis. Interestingly it is thought to reach kidney cells through binding (less preferably) to Gb4 which is present on epithelia endothelial cells (Betz et al. 2011; Ho et al. 2013).

Stx is divided into two main subtypes, Stx1 which heavily resembles the originally isolated Stx within *Shigella dysenteriae*, and Stx2 which only exhibits approximately half the sequence identity with the original Stx1 (Kaper, Nataro, and Mobley 2004). These can be further subdivided into subtypes: Stx1a, Stx1c, Stx1d, Stx2a, Stx2b, Stx2c, Stx2d, Stx2e, Stx2f, and Stx2g. In locations where EHEC O157 strains are a threat to human health, the most severe disease is associated with strains encoding Stx2a (Dallman et al. 2015). Somewhat confusingly these Stx subtypes are composed of two protein subunits named subunit A and subunit B. Subunit B being the part of the toxin which binds to Gb3 (Fuller et al. 2011). It is notable that Stx2a and Stx2c only exhibit six amino acid differences within their protein sequence even though Stx2a exhibits a higher level of pathogenicity compared to other subtypes including Stx2c (Fuller et al. 2011). It should be noted that these two Stx subtypes are the most associated with serotype O157:H7 (Dallman

et al. 2015), thus potentially explaining why this serotype is responsible for such a relatively large proportion of EHEC human infection.

The comparison of different *E. coli* strains, in particular an EHEC O157 isolate and an *E. coli* K-12 isolate (commensal) provided clear evidence that Stx genes were encoded on bacteriophages (O'Brien et al. 1984) that were integrated into the *E. coli* genome as prophages.

1.3 Bacteriophages and Prophages

Bacteriophages are viruses which only infect bacteria (Ofir and Sorek 2018). Bacteriophages act in the same way as human viral infections in that their genetic content enters the bacterial cell and utilises the cellular machinery to replicate and propagate its genetic content (Ofir and Sorek 2018). For a sub-set of bacteriophages, prophages are a by-product of this process, where the bacteriophage genome integrates into the bacterial chromosome. The viral genome is replicated along with the bacterial genome (Ofir and Sorek 2018). Standard 'lytic cycle' replication is resumed when the bacterium is faced with specific stimuli (mostly damaging ones) in which the prophage is induced and the phage genome excises, replicates and the copies are packaged into nascent phage particles before release via bacterial lysis (Asadulghani et al. 2009). This mechanism is not only ingenious but fascinating. If we perceive the goal of all "living" things (viruses are not technically living) to propagate their genetic content, this mechanism offers many advantages. Firstly, the prophage normally propagates its content as part of the bacterial chromosome as the bacterium naturally replicates. However, the moment a stress stimulus is received, it causes the bacterium to propagate it further in order to potentially "infect" other bacteria.

For EHEC infections other by-products of this interaction are critical, in particular Stx which is expressed and released during prophage induction and bacteriophage release respectively (Balasubramanian et al. 2019). The fact that Stx were encoded by a prophage was determined long ago (O'Brien et al. 1984), however, many key studies further investigated the EHEC O157 genome and its horizontally acquired genetic material. These horizontally acquired genes were first term O islands (OIs) and were relatively more common in the EHEC O157 genome than other *E. coli* genomes studied at the time (H. Schmidt and Hensel 2004; Imamovic et al. 2010). Within these OIs were Pathogenicity Islands (PAIs) and prophages. The main difference between PAIs and prophages is the fact that prophages are their own entity, "aiming" to propagate its own genetic content first. As mentioned Stx-

encoding prophages are highly important to the pathogenicity of EHEC O157. However, other non-prophage encoded factors such as the Locus of Enterocyte Effacement (LEE) PAI are also highly relevant pathogenicity factor (Mcdaniel et al. 1995). The LEE encodes the intimin (*eae*) gene previously mentioned, and is a key part to the attachment of the bacteria to the terminal rectum of the host (Mcdaniel et al. 1995; Dallman et al. 2015). Furthermore PAIs as well as prophages are able to affect the regulation of other genetic material (H. Schmidt and Hensel 2004; Tobe et al. 2006; Asadulghani et al. 2009; Ogura et al. 2015). However, this body of works focuses on prophages rather than all OIs as these appear to be more “active” and “consistent” as a population (while offering a great level of variety) than PAIs.

While bacteriophage gene content can be small, they can also carry genes that can impact the bacterium’s fitness, ability to survive, and pathogenicity greatly (Ooka et al. 2009; Dallman et al. 2015). Therefore, bacteriophages are a key part of horizontal gene transfer. However, they also offer a potential pathway to kill bacterial cells and have, therefore, been greatly studied as alternatives to antimicrobial drugs (phage therapy). However, this field is still in its infancy as much is yet to be understood about the mechanism behind which phage will lyse which bacteria (phage therapy targeting). Furthermore, there appears to be a point where prophages lose their identity as prophages and become integral parts of the bacterial chromosome, meaning that these prophages become unable to induce or lyse the cell (Asadulghani et al. 2009). All these phenomena have been well studied in EHEC O157, as well, as its prophage content, and these topics will be further discussed in **Chapters 2** and **3**.

1.4 Classical Typing Methods of EHEC

In the UK the predominant serotype associated with human infections is O157:H7 (Chase-Topping et al. 2008). Two main Public Health (PH) bodies oversee the tracking of EHEC outbreaks in the UK. These are Public Health England (PHE) (working jointly with Public Health Wales (PHW) and the Public Health Agency (PHA) (Northern Ireland)), and the Scottish *E. coli* Reference Laboratory (SERL) (works jointly with Health Protection Scotland (HPS)). A key aim of these PH institutions is to type and define EHEC clinical (and occasionally veterinary) isolates, as well as to help trace the origin of an outbreak to facilitate and inform possible interventions.

To track outbreaks multiple typing methods were developed and utilised. These included but were not limited to: serotyping (Blanco et al. 1996), phage typing (Ahmed et al. 1987; Khakhria, Duck, and Lior 1990), Pulse Field Gel Electrophoresis (PFGE) (Izumiya et al. 1997; Byrne et al. 2014), Multi-Locus Sequence Typing (MLST) (Afset et al. 2008), Multi-Locus Variable-number tandem repeat Analysis (MLVA) (Pei et al. 2008; Byrne et al. 2014), and Shiga Toxin (Stx) typing (Scheutz et al. 2012). Some of these methods heavily relied on DNA amplification and targeted specific parts of the genome .

Serotyping is the process by which a species of bacteria is subclassified based on surface antigens. This simple form of typing relies on antibody detection of surface variants such as in Lipopolysaccharide (LPS) and flagellin and therefore is of relatively low resolution. However, it still offers valuable insights as certain serotypes are associated with higher risk of diseases than others, most notably EHEC O157:H7 (Ratnam et al. 1988). This serotype was first detected in 1982, and involved distinguishing the O157 and H7 antisera (Ratnam et al. 1988). Nowdays it is possible to determine the serotype of most EHEC isolates based on sequences of the relevant chromosomal regions and Polymerase Chain Reactions (PCRs) (Fratamico et al. 1995; Feng and Monday 2000).

Phage typing, is a method which classifies strains based on their sucseptibility or resistance to a panel of lytic bacteriophages. This panel was first developed in Canada in 1987 (Ahmed et al. 1987) before being extended in 1990 (Khakhria, Duck, and Lior 1990). This method yields interesting insight into the EHEC O157 isolate population, with certain phage types more common in human infections (Chase-Topping et al. 2008; Dallman et al. 2015). However, the genetic differences underlying variation in phage susceptibility generally (and also in the case of EHEC) are still not fully understood. As we will explore further on in this thesis, there are relationships between the prophage content of EHEC isolates and phage resistance and susceptibility phenotypes: as the prophage content can change so can the phage typing results (Cowley et al. 2016).

PFGE typing is a method which uses endonuclease digestion, and agarose gel electrophoresis to first cut up and then separate the genome into different sized DNA fragments, and then subclassify isolates based on the banding patterns. In the case of EHEC the digestion site typically used is that targeted by *XbaI*, and papers can be found using this method for outbreak tracking (Izumiya et al. 1997). However, it should be noted that this method has many limitations. For one it is a quite difficult and time-consuming protocol with multiple steps (Lesley Allison, SERL, Personal communication). In the analysis stages the reading of the results is quite subjective, even when using the same equipment and settings (Keim et al. 2000). Therefore, when tracking outbreaks nationally or internationally, it is a challenge to standardise the protocol and analysis of results. However, this was achieved and PFGE was and still is a major epidemiological tool for EHEC studies with large databases such as PulseNet (Ribot et al. 2006).

MLST typing started becoming popular in the late 1990s. This method requires a panel of housekeeping or highly essential genes to be selected (Maiden et al. 1998; Maiden 2006; Belén, Pavón, and Maiden 2009). A region of 500-600 bp of these genes is sequenced and the different allelic variants are defined. The allelic composition of each strain is determined (for

these seven genes), and every different combination is given a unique identifier known as a Sequence Type (ST) (Maiden 2006; Belén, Pavón, and Maiden 2009). This method was and is succesfully used in evolutionary and epidemiological studies (Afset et al. 2008; Belén, Pavón, and Maiden 2009) as it truly types genome variation, especially Single Nucleotide Polymorphisms (SNPs), albeit in a limited number of genes.

Finally, late in the first decade of the 2000s, MLVA typing was developed. This method relies on loci that had been previously studied with their mutation rate and diversity known. These loci are amplified by Polymerase Chain Reaction (PCR), analysed by gel electrophoresis, which gives a size estimate for the loci, which in turn could be used to predict the number of repeats in each locus (Nadon et al. 2013). These patterns are classified and serve as identifiers for different versions of the loci. The EHEC O157 MLVA panel originally involved nine loci, but has since been extended to be usable for other EHECs (Izumiya et al. 2010).

Another typing method which aims more to predict phenotype and health risk rather than clustering is Stx typing. This is a PCR-based method which detects the presence and subtypes of the *stx* genes (Scheutz et al. 2012). However, this is not a straightforward endeavour. The two most common Stx subtype in the UK are Stx-2a and -2c. Therefore, most Stx typing would likely aim to detect and differentiate between these, especially considering that Stx-2a tends to be associated with more severe symptoms than Stx-2c (Persson et al. 2007). However, there are only 22 bp different between the two subtypes, resulting in only a six amino acid difference (of these six amino acid differences, only two do not exhibit any property preservation), thus making it hard to differentiate these at a PCR level.

1.5 DNA Sequencing

DNA sequencing is the process by which the order and identity of nucleotide bases within a strand of DNA are determined. In this thesis this process will often be referred to as Whole Genome Sequencing (WGS). This is a slight misnomer as WGS can refer to any type of DNA sequencing that covers the whole genome of the organism of interest. However, recently WGS is synonymous to Next Generation Sequencing (NGS). Ironically, NGS is also a misnomer as the generation of sequencing technologies it refers to has already been implemented and heavily improved. Therefore, NGS tends to refer to any type of high throughput sequencing, of which there are three main successful platforms: Illumina, Pacific Biosciences (PacBio), and Oxford Nanopore (Update 2019: Illumina is in the process of acquiring PacBio). Also, I will be using WGS to refer mostly to bacterial WGS.

When looking at NGS there are five key factors to take into consideration: throughput, random error rate, read length, turn-around time, and cost. Platform size, portability, and library preparation protocol are other important factors for certain use case scenarios, but for this project, these are unimportant.

Illumina (previously known as Solexa) (Illumina 2019) is one of the more widely used platforms (specifically the MiSeq sequencer), especially in PH microbiology. This is due to its relatively low cost for a high throughput (dependent on the chemistry chosen, the following will only focus on the more recent 150, 250, and 300bp chemistries) (£60 per bacterial isolate for a coverage of ~40x) with a low random error rate (< 0.01%). Illumina's weakness, compared to the other two technologies mentioned, is that its read length ranges from 50-300 base pairs (bps) ("short-read" sequencing). This is due to the chemistry behind Illumina sequencing, called Sequencing By Synthesis (SBS) which uses a PCR amplification step. This method requires the DNA content to be amplified so that a single DNA molecule becomes a cluster. This cluster of identical DNA molecules is then sequenced as

nucleotide bases are added with fluorescent tags. If the DNA molecules are synthesised in phase, they output the same fluorescent signal at each cycle. However, as time proceeds the DNA molecule starts synthesising out of sync, which results in a muddled signal. This is the reason the quality of Illumina reads tends to drop as the read gets longer towards the end of the read. Therefore, SBS currently has a maximum read length it cannot go over and it also can be affected by PCR amplification bias (Aird et al. 2011).

The PacBio and Oxford Nanopore platforms do not use PCR amplification or a signal that requires phasing that can be lost as sequencing progresses and therefore can generate “long-read” sequences. They both use “single-molecule” technologies. The PacBio platforms use Single-Molecule Real-Time sequencing (SMRT) (Pacific Biosciences 2019). SMRT sequencing utilises zero-mode waveguides (ZMW), which are wells within the flowcell at the bottom of which a single DNA molecule can be attached. Due to the properties of that ZMW the light emission of single base addition events can be detected (it utilises nucleotide bases with different colour fluorescent markers as does Illumina). Therefore, as the signal doesn’t degrade in the same manner, SMRT sequencing offers reads typically ranging between 10 and 15 kbp. SMRT sequencing offers another advantage, as the signal observed is from a single DNA molecule, the speed of the process can be monitored. A normal DNA polymerase will synthesise three bases per second. However, if a DNA base is methylated (methyl groups attached to the nucleotide base), the process slows done. As such, PacBio sequencing can offer a methylation profile as well as a DNA sequence, which can be used to investigate the epigenetics of the organism being sequenced (Ardui et al. 2018; Pacific Biosciences 2019). While these are all advantages that the Illumina platforms do not possess, PacBio is disadvantaged due to its relatively high error rate (14%) (Ardui et al. 2018); relatively low throughput (the throughput is now comparable to that of a MiSeq, however, higher coverage is required to account for the higher error rate), which translates into a higher overall cost per bacterial sample (£200-£500 at time of project start) (Jim Bono, USDA, Personal communication). This, coupled with a

rather large upfront capital cost, makes PacBio less suited for PH investigations and more suited for an academic setting.

Oxford Nanopore's MinIon is the latest offering in terms of sequencing technologies. It also obtains sequencing reads for single DNA molecules, and therefore doesn't suffer from any of the drawbacks associated with Illumina sequencing. Unlike PacBio and Illumina, the MinIon doesn't use SBS. It generates a sequence as the DNA strand passes through the nanopore, a protein set in an electrically resistant polymer membrane, with an ionic current passing through it (Oxford Nanopore 2019). As DNA bases pass through the pore, the current of it changes, and that change in current is monitored and translated into a DNA sequence. This means that the limit for MinIon read length is the lifetime of the nanopore, which is currently around 48 hours of sequencing (reads up to 1 Mbp have been recorded). It is, however, difficult to quantify the throughput possible within these 48 hours. While, MinIon advertised around nine GB of throughput per run in 2016 (Oxford Nanopore 2019), it has been noticed that throughput is highly dependant on read length targeted. The longer the reads, the fewer DNA molecule ends are in the solution, the less chances one passes through a pore (for example if the same concentration of DNA is sheared into bits of 3000bp or bits of 1 Gbp, the former will have a higher level of chance that a DNA end goes through a pore and starts the sequencing process). In addition, this doesn't consider the relatively higher level of difficulty in creating a sequencing library with long reads. Finally, due to the organic aspect of the MinIon (organic nanopores) which gives flowcells a shelf life, a more in-depth experiment design needs to be created, as to not waste any of the available throughput (for example by washing flowcells if not fully utilising its sequencing potential in a single run). This coupled with the still relatively high error rate (~3-8%), and relatively high price per bacterial isolate if a decent read depth is wanted (~£200) (David Greig, PHE, Personal communication), makes it better suited currently for academia rather than PH.

To this day, the best overall approach remains a hybrid run between Illumina and a “long-read” sequencing technique, to obtain a fully closed genome with high levels of confidence for each base called. Interestingly, recent news has been released about Illumina acquiring the PacBio company, therefore, an official hybrid sequencing platform might already be in the works. However, this will obviously come at a cost, a cost which may be much lower than expected as Oxford Nanopore’s MinIon keeps evolving (Flongle, SmidgIon, etc), and running it keeps getting cheaper. **Table 1.1** regroups the advantages and disadvantages of all three platforms discussed (for Illumina platforms the focus is on the MiSeq platform as it was, at the time of this project, the most common, sequencing platform in the PH field).

Platform	Technology Type	Accuracy (%)	Throughput (Gb)	Cost per isolate (£)*	Run Time (hrs)
Illumina (MiSeq)	Short-read (150 – 300 bp)	99.9	4 – 15	60	24 - 56
PacBio (Sequel)	Long-read (10 – 15kbp)	87	5 – 10	200*	4
Oxford Nanopore MinIon	Long-read (3 – 1000 kbp)	93 - 97	6 – 9**	95 - 200*/**	48

Table 1.1 Table regrouping the features of each sequencing platform. This table utilises the data for the PacBio sequel, while as later described, the PacBio RSII was used for this project. However, the numbers for the later chemistries of the RSII are closer to those of the sequel than those it was originally advertised with.
* Given an optimised workflow, and not considering staffing cost.
** Highly dependent on targeted read length.

1.6 Genomics, Phylogenetics, and Phylogenomics

Once the DNA sequence of the isolates of interest are obtained, different types of analyses can be conducted. The types of studies conducted in this thesis can be divided into two categories: genomics and phylogenetics. Genomics is the field of biology focusing on genomes, which are the complete genetic content of an organism, in other words its whole DNA sequence, or whole genome sequence. Phylogenetics is the field focusing on the evolution of an organism but can use many different aspects from biology to achieve this goal. Phylogenomics is a subset of phylogenetics merged with genomics, where evolutionary relationship is inferred through the genomic content of an organism. This thesis will be mainly a genomics study, where many different aspects of the organism's genetic content are analysed in order to determine various features of the organism in question. However, **Chapters 2**, and **3** will include phylogenomics analyses. As such, certain key concepts need to be introduced: pangenome, core genome, accessory genome, gene expression, and methylation.

When discussing genomics, it is important to define terms clearly, as certain individuals tend to use the terms genomics and genetics interchangeably. Genetics technically refers to the DNA content of an organism in the forms of genes, it therefore, does not consider non-coding DNA regions. Genomics, on the other hand, represents the complete DNA content of an isolate. The core genome is the genomic content that is present across all strains of a species (Xiao et al. 2015). Even though when referring to the core, accessory, and pangenome, they are referred to in comparison to a species, in most cases, they are used towards a specific sequence set of the species of interest, as it would be quite challenging to sequence all the variation present in an actual species. However, the core genome can be defined differently. Terms such as soft core are becoming even more common as scientists investigate different presence thresholds to define core genomes. Accessory genome on the other hand is the genomic content that is present in only a subset of isolates of a species (Xiao et al. 2015). Once more, the

exact definition of this can change with certain individuals setting the accessory genome to any genomic content present within even a single isolate, while others will prefer it to be present in at least a specific percentage of the species. Putting together the core and accessory genome yields the pangenome: the complete genomic content of a species (Xiao et al. 2015). Each of these different aspects of genomics play a role in different analyses.

The core genome is typically used for SNP phylogenies (which may be a misnomer considering phylogenies are also called phylogenetic analysis, but newer phylogenies also consider the non-coding DNA sequences). SNP phylogenies aim to determine how related isolates are by how many mutation points differ between them. This is based on the logic that a SNP is at its most basic a single nucleotide mutation point. Assuming a steady mutation rate across a species, or clade, it would therefore be logical to assume that the number of mutation points is directly relevant to the relatedness between isolates. However, caveats exist: certain areas of the genome can have higher or slower mutation rates, an isolate can become hypermutable, and environmental factors can influence mutation rate (S. Schmidt et al. 2008; Maharjan and Ferenci 2018; Sharp et al. 1989). However, even with the above caveats core-SNP phylogenies have been proven to be highly reliable (Eppinger et al. 2011; Underwood et al. 2013; Cunningham et al. 2017; Holmes et al. 2018). However, the accessory genome cannot be used for this type of analysis as it would introduce a high number of SNPs which were not necessarily obtained through mutation but horizontal gene transfer. The same logic is applied for areas with high levels of recombination.

The accessory genome can be of use in many different types of studies such as looking at virulence factors or antimicrobial resistance which typically have mechanisms that involve horizontal gene transfer. However, in this study I will discuss how the accessory genome can also be useful to supplement phylogeny data. In this thesis I will be focusing the majority of **Chapters 2** and **3** on prophages (**Section 1.1**) which tend to be accessory genome,

except for a few which are present in most isolates studied and were therefore most likely acquired a long time ago. In the case of EHEC O157 the accessory genome is more important than usual due to the high number of prophages present within the genome, and the presence of key virulence factors within these prophages.

The pangenome is typically used to study the presence and absence of genes, and inferring relatedness, as well as potential phenotypic differences between strains (Xiao et al. 2015). However, depending on the sequencing technology used different aspects of genetics can be explored. For example, if sequencing RNA molecules (through a process called RNA-seq), one can determine whether genes are differentially expressed within different samples. This allows for experiments to determine the effects that the environment or gene editing can have downstream. For example, one could use RNA-seq to determine whether a structural chromosomal variant results in genes having differential expression. Another way to investigate similar questions is through PacBio sequencing. As previously mentioned, PacBio sequencing allows for the methylation profile of DNA molecules to be determined, this can be used to study the epigenetics of an organism and once more how certain factors can affect gene expression. All these analyses typically do include the accessory or pangenome.

The clonal origin of EHEC O157 allows for the proposed analysis of its accessory genome to be more feasible., Interestingly, *E. coli* as a species appears to only have 20% of its genome as core and, therefore, leads the question of whether it should be a single species.

1.7 Supplementation of EHEC Typing Using Whole Genome Sequencing

A point to note is that all the previously discussed typing methods (**Section 1.4)** were not meant to replace one another through time. These were methods that clustered and categorized isolates based on different aspects and could be used in juxtaposition with one another. Therefore, the advent of WGS was truly the start of a new era. WGS had the potential to replace these methods and offer very high-resolution clustering.

EHEC typing was first supplemented by “simple” nucleotide sequencing. These techniques, such as MLST, focus on sequencing a specific part or parts of the genome, and clustering strains based on that part. Specific area sequencing is still used in PH when looking at certain viruses, such as HIV. However, due to the typically higher complexity and genome size of bacterial pathogens, partial sequencing has its limitations. As technologies evolved, sequencing throughput started rising dramatically. It is now possible to obtain 120 Giga-base pairs (Gbp) of short-read sequencing data in a single 30 hours run, using the newest Illumina platform (Illumina 2019). This allows for WGS and high coverage sequencing of many bacterial isolates, thus permitting the tracking of clinically relevant strains within the PH system. One should note though that WGS does not mean that fully assembled closed genomes are made available, but that the whole genome is used for the DNA extraction and preparation (as opposed to targeted sequencing). To obtain fully assembled genomes different sequencing techniques need to be deployed, and these will be discussed in due course. However the data obtained from short-read WGS can already greatly supplement, if not lead, outbreak tracking.

The primary goal of WGS in PH is to answer four questions. The first one is the identity of the pathogen, the second one is its origin and relationship to other isolates, followed by its prevalence in the public, and finally the dangers it poses. Many different methods have been and are being developed to

obtain this information, some of which will be discussed further on. In the UK, one of the key workflows used for EHEC O157 WGS is based on the tool SnapperDB developed at PHE (Dallman et al. 2018). An instance of this tool has also been deployed at the Scottish *E. coli* Reference Laboratories (SERL) to allow for a UK wide approach to EHEC O157 tracking (Holmes et al. 2018) (PHW samples are ran by PHE on the SnapperDB workflow (personal communication with Dr. Tim Dallman)).

The SnapperDB workflow is a SNP phylogeny analysis which first requires the reads to be aligned to a reference sequence (for all EHEC O157:H7 the reference sequence was the EHEC O157:H7 Sakai strain, accession number: BA000007), and if bases are covered at a depth of 15x and differed from the reference one, they are recorded as SNPs. This is done for any strain of interest. However, to compare these strains, only regions of the genome that are mapped by more than 95% of the strains within the tested population are kept. This is the core genome and including the rest of the genomic content (the accessory genome) could highly skew the results without providing any phylogenetically relevant data. However, this is only due to the fragmentation of the accessory genome when it is sequenced using short-read sequences (discussed in **Chapter 2**). Variant calling used the PHE tool PHEnix (PHE Bioinformatics Unit 2015), which utilizes BWA (Li and Durbin 2009) for the read mapping, and GATK for SNP calling (McKenna et al. 2010).

Once core SNPs are determined, the PHE software SnapperDB is then applied (Dallman et al. 2018). The aim of this program is to cluster all the given strains based on their SNP distance. It does so through a SQL database and single-linkage clustering. First SnapperDB calculates pairwise SNP distances: this is the distance between any two strains and is the number of SNPs that differ between the two. Once all pairwise distances are calculated and stored (to avoid having to be repeated) in the SQL database, a distance threshold is determined, and the clusters generated. The SNP distance thresholds tested are as follow: 250.100.50.25.10.5.0. It should be

noted that one of the key features of SnapperDB is the nomenclature of its clustering. It is a seven-digit code separated by dots (e.g. 4.4.4.4.28.54.215). Each of these digits represents a unique cluster identifier for a specific distance threshold. In other words, the first digit of the code is the cluster identifier for a distance threshold of 250 SNPs, the second digit for a distance threshold of 100 SNPs, and so on.

What makes this nomenclature powerful, is the ability to quickly understand the relationship level between any two strains. However, while one may logically assume that two strains within the same 100 SNP cluster identifier must have 100 or less SNPs between them, that is not the case. This is because single-linkage clustering is used. This means that if a strain is within 100 SNPs of even a single strain within the cluster, it will be integrated as part of the cluster. However, to stop excessive cluster merges (merges between clusters occur when an isolate is found to be within the distance threshold of members of two different clusters), an outlier threshold is in place. If an isolate is found to belong to a cluster, but the average of its SNP distance to the other members is more than two standard deviations, from the averages of the other members to one another, an outlier flag is raised. This indicates that the user should manually curate the sequence deciding whether to include or ignore it. One key drawback of this method is the difficulty in standardising the output across multiple sites independently.

1.8 The Future of WGS in Public Health

WGS is currently moving in strides and greatly changing PH. The main advantage being a single cost, which allows for multiple typing tests at a much higher resolution. It is common for PH labs to run multiple typing tests such as MLVA typing and MLST to be able to compare data with other labs, as well as having to run new tests on older samples. WGS offers the advantage of providing that data in a way that if new tests are developed the sequencing data is still available for analysis. However, PH has limitations, one being return on investment, the second being a need for validated and accredited protocols, and the third is the need for rapid turnaround times. WGS capabilities first require a relatively large investment which is then followed by regular machine maintenance, reagent, and data analysis costs. While this does not differ from other typing techniques, WGS is evolving fast and platforms can quickly become obsolete.

Secondly is the geopolitical aspect of PH. Worldwide PH is monitored partially by the Centers for Disease Control and Prevention (CDC) in the USA, and the European CDC (ECDC) in Europe. This means that for worldwide outbreak tracking, costs need to be taken into consideration and high costs could limit the involvement of many low- and middle-income countries. Financial limitations prevent purchasing of newer platforms, and expensive compute and software. Attempting to standardise data analysis between all these countries is a challenging endeavour, which is in part why PH labs cannot reach the forefront of scientific development. They require methods that have been heavily tested and provide consistent results. Considering how long WGS has been available in the scientific field, and that it is only now becoming the norm in PH is proof of that. Therefore, currently the leading platform in PH is Illumina sequencing. It offers high throughput at a low cost consistently. However, using newer, less optimised platforms (at the time of this project), such as the Oxford Nanopore MinIon (Oxford Nanopore 2019), could reduce this cost and generate relevant data that cannot be generated using short-read sequencing (as we will see in further

chapters). However, these platforms are not nearly as consistent, require more specialised individuals, and lead to the question: is more data required?

While scientific curiosity dictates that more data is better, it does not always allow for questions to be answered, as much as leading to new questions. Even more so, it is quite possible that the data currently generated from PH labs is all that is required for current health protection. There is very little debate of what WGS brought to PH: allowing for finer typing, following a near identical protocol for a variety of organisms, and allowing for different methods of typing in one procedure while also being simpler to standardise. However, it could be argued that current health protection agencies do not require more specific data. Using the data currently generated, one can trace outbreaks, determine common ancestors between strains, and how related strains between individuals are. This enables us to quickly determine whether a case is sporadic, or part of an outbreak, and then potentially trace the origin of the outbreak. With current data, the bottleneck is not typing resolution anymore, but metadata obtained from the infected individuals and data sharing. Novel questionnaires need to be designed to truly allow health protection agencies to gather all the required travel and interaction information. However, results will still be limited if no global database of typing results is made available. An outbreak strain could be found having originated from an individual that has travelled, however only with typing data from the population of the country of travel will the source of the infection become clear. This requires coordination between international PH labs, which brings us back to the CDC and ECDC.

These institutions goals are to coordinate data between PH labs, and as such determine the typing tests to be conducted on isolates. Currently these organisations agree for the need of WGS and are making a push for core genome MLST (cg-MLST). Cg-MLST is a newer WGS typing method, which unlike SNP based typing (such as SnapperDB) revolves around determining the allelic compositions of genomes. As opposed to normal MLST, it does not use a seven gene scheme, but utilises all genes that compose the core

genome of a species. This allows for typing schemes using hundreds, if not thousands of genes. Many papers in the literature (Cunningham et al. 2017; Janowicz et al. 2018; Pearce et al. 2018) compare cg-MLST and SNP typing methods and find little differences in the generated typing results. SNP typing does offer a higher resolution, however this resolution does not appear to be needed for PH and may be more relevant in research areas. Furthermore, cg-MLST requires less compute and is more easily scalable than SNP typing. One could even argue that such large cg-MLST schemes are not necessary, and that using a 300 gene scheme might already provide enough resolution while lowering compute cost. However, the challenges of cg-MLST are not in the compute, but in the generation and maintenance of the scheme. This is currently the reason why; since the CDC and ECDC have implied their backing of cg-MLST, no larger steps have yet been taken. Cg-MLST requires a centralised entity to take control of the scheme and curate novel alleles as they are detected for each species. This is obviously a challenging task with limited funding potentials. This uncertainty and delay are why SERL in parallel with the SnapperDB pipeline, also run cg-MLST. This allows them to have a UK wide typing method already deployed, while being prepared for the deployment of worldwide cg-MLST when it occurs.

This means that the PH field is currently in a period of great flux, which will most likely emerge in the WGS sequencing era. However, while we did mention that more data might not always be beneficial, it should be noted that this is for the current needs of PH. I believe that once the field of metagenomics is more fully developed, there will be a need for this type of data to be generated, which will cause another flux as metagenome data is extremely hard to assemble using short-sequencing reads. However potentially, prior to this flux, might emerge the need for long-read sequencing data as these platforms allow for full genomes to be assembled in single contigs, and this offers a novel insight in pathogen strains as we will see in further chapters.

1.9 Project Aims

The main aim of this project was to use bioinformatics analyses to improve the diagnostic pipeline of EHEC, and its real-time-epidemiology. To achieve this the project was subdivided into two key areas:

- Using the advantage offered by PacBio sequencing to understand variation in the accessory genome, especially prophages, in human and bovine *E. coli* O157. **Chapters 2, 3, 4**.
- Applying the knowledge obtained from long-read sequencing data and analyzing how it can supplement public health data. **Chapters 2, 3, 4**.

1.10

2 First Look at the Prophage Population of EHEC O157:H7 and a Genetic Content Analysis

2.1 Preface

2.1.1 Research Overview

The previous chapter introduced the typical analysis of the core genome of EHEC O157:H7 for health protection applications. When this PhD project first started (October 2014), the accessory genome, the genomic content that is not shared by all EHEC O157:H7, was mainly disregarded due to the difficulties assembling it using short-read sequencing technologies. Much of the literature available at the time investigated the core genome, or very specific areas of the accessory genome such as the Stx encoding regions (Underwood et al. 2013; Ashton et al. 2015). However, with the decrease in long-read sequencing costs, and the advent of the Oxford Nanopore platform, more papers are being published investigating the accessory content of EHEC O157:H7 (Saile et al. 2016; González-Escalona et al. 2019). One such paper was written by myself, members of the David Gally group, and other collaborators, and is the core of this chapter. At the time of the paper nine UK strains were sequenced using the Pacific Biosciences (PacBio) long-read sequencing platform. The reads generated were able to fully assemble and close the genomes, which were then used to identify, extract, and analyse the prophage regions present in these strains.

This paper was a first look at the prophage population within EHEC O157:H7 in UK strains from cattle and human sources, which were also selected to represent the diversity found in UK isolates. This was conducted with the aim of investigating the variation present in prophages within the UK population of EHEC O157:H7 strains, as well as to identify genomic information that could be used to better understand the molecular biology of the species. As covered in the later discussion, the analysis presented was extremely limited by the low overall number of sequences used in the different analyses. However, I consider the data presented in the published paper was an

important stepping stone to understand the evolution of this important zoonotic pathogen and raised many intriguing points regarding the variable prophage population in the *E. coli* species, including deliberate maintenance of regions of prophage homology to promote large chromosomal re-arrangements. These will be analysed and discussed in future chapters.

Due to the format restrictions on published papers, only a limited Introduction and Methods section is presented. These will be explained in much more detail in the next chapter (**Chapter 3**), where I will go into the particularities, mechanisms, and rationale for use of the tools utilized, as well as the iterations this analysis underwent to reach this specific published version, as this represents a stepping stone to the final analysis included in this thesis. Supplementary material can be found in **Appendix I**.

2.1.2 Declaration of Own Work

The work presented in this paper was conducted by myself with the help of the listed authors. PacBio sequencing and genome sequence assembly was conducted by Dr. Jim Bono. Strain selection, and UK wide phylogeny was conducted by Dr. Tim Dallman (PHE), with clinical isolates being obtained from PHE and SERL. Cattle isolates were obtained through the IPRAVE project, a Wellcome Trust-funded epidemiology programme that involved collection and analysis of *E. coli* O157 from cattle in Scotland between 2002-2003. Laboratory work and analysis regarding prophage integration, and phage typing was conducted by Dr. Lauren Cowley and Sean McAteer. Supervision of the project and help with the writing of the manuscript was provided by Prof. David Gally. The complete Bioinformatic analysis of the long-read generated genomes, and their prophage content was conducted by me.

2.1.3 Strain Selection

The work presented in the manuscript below centers around 14 sequences. Five of these sequences were previously published sequences, four of which were chosen due to being widely used and accepted reference sequences in the literature:

- Sakai (accession number: BA000007) is one of the first sequences of EHEC O157 that was made available. While it is the sequence of a Japanese strain, it has been widely accepted and used as the primary sequence reference for EHEC O157.
- SS52 (accession number: NZ_CP010304) was, at the time of this project, a new sequence of a USA super-shedder strain obtained through long-read sequencing (Katani et al. 2016). Considering, the interest of the IPRAVE project with super-shedding, and the rarity of long-read sequenced genomes at the time of this project, this sequence was considered a relevant reference sequence to be used.
- TW14359 (accession number: NC_013008) is a widely used reference sequence from a strain that was related to a large food-borne outbreak in the USA (Kulasekara et al. 2009).
- EC4115 (accession number: NC_011353) is a sequence of a strain related to the same food-borne outbreak as TW14359 (Eppinger et al. 2011).
- EDL 933 (accession number: NC_002655) is a sequence from a food-borne outbreak in the USA isolate in 1982. It has since, been widely used as a reference strain for EHEC O157 (Perna et al. 2001).

The remaining nine sequences used in this chapter were generated specifically for this work. Strain selection for these nine sequences was based around previous work conducting by Dr. Dallman (Dallman et al. 2015) and the IPRAVE project. The aim was to select for an even representation of UK isolates across the phylogeny run by Dr. Dallman (Fig. 1 in the paper

below). This involved having a diverse selection of lineages and PTs as described in the manuscript below.

2.1.4 Code availability and Versioning

All code written for the purpose of this chapter can be found at: https://github.com/SharifShaaban/PROPI. The following tool versions are required: Python 2.6.6 and 2.7.9 (required modules: os.path, sys, subprocess, shutil), EMBOSS 6.5.7.0, BLAST+ 2.2.28, R 3.0.0 abd R 3.2.2 (required libraries: magrittr, readr, ggplot2, cowplot), Perl 5.18.1, Prokka 1.5.2 (Prokka requires its own dependencies), Get_Homologues 1.0 (Get_Homologues requires its own dependencies), and Easyfig CL 2.1 (Easy_fig requires its own dependencies).

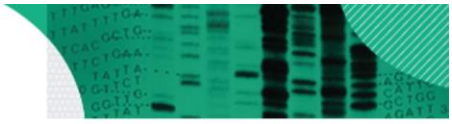
2.1.5 Paper reference

(Shaaban et al. 2016)

S. Shaaban, L. A. Cowley, S. P. McAteer, C. Jenkins, T. J. Dallman, J. L. Bono, D. L. Gally. Evolution of a zoonotic pathogen: investigating prophage diversity in enterohaemorrhagic *Escherichia coli* O157 by long-read sequencing. *Mgen*, 2016

2.2 Manuscript

MICROBIAL GENOMICS
Bases to Biology



Research Paper

Evolution of a zoonotic pathogen: investigating prophage diversity in enterohaemorrhagic *Escherichia coli* O157 by long-read sequencing

Sharif Shaaban,[1] Lauren A. Cowley,[2] Sean P. McAteer,[1] Claire Jenkins,[2] Timothy J. Dallman,[2] James L. Bono[3] and David L. Gally[1]

[1]Division of Infection and Immunity, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush EH25 9RG, UK

[2]Gastrointestinal Bacterial Reference Unit, 61 Colindale Avenue, Public Health England, London NW9 5EQ, UK

[3]U.S. Meat Animal Research Center, Agricultural Research Service, U.S. Department of Agriculture, Clay Center, NE 68933-0166, USA

Correspondence: David L. Gally (dgally@ed.ac.uk)

DOI: 10.1099/mgen.0.000096

Enterohaemorrhagic *Escherichia coli* (EHEC) O157 is a zoonotic pathogen for which colonization of cattle and virulence in humans is associated with multiple horizontally acquired genes, the majority present in active or cryptic prophages. Our understanding of the evolution and phylogeny of EHEC O157 continues to develop primarily based on core genome analyses; however, such short-read sequences have limited value for the analysis of prophage content and its chromosomal location. In this study, we applied Single Molecule Real Time (SMRT) sequencing, using the Pacific Biosciences long-read sequencing platform, to isolates selected from the main sub-clusters of this clonal group. Prophage regions were extracted from these sequences and from published reference strains. Genome position and prophage diversity were analysed along with genetic content. Prophages could be assigned to clusters, with smaller prophages generally exhibiting less diversity and preferential loss of structural genes. Prophages encoding Shiga toxin (Stx) 2a and Stx1a were the most diverse, and more variable compared to prophages encoding Stx2c, further supporting the hypothesis that Stx2c-prophage integration was ancestral to acquisition of other Stx types. The concept that phage type (PT) 21/28 (Stx2a+, Stx2c+) strains evolved from PT32 (Stx2c+) was supported by analysis of strains with excised Stx-encoding prophages. Insertion sequence elements were over-represented in prophage sequences compared to the rest of the genome, showing integration in key genes such as *stx* and an excisionase, the latter potentially acting to capture the bacteriophage into the genome. Prophage profiling should allow more accurate prediction of the pathogenic potential of isolates.

Keywords: Shiga toxin; bacteriophage; prophage; *Eschrichia coli* 0157.

Abbreviations: diya, Do-It-Yourself Annotator; EHEC, enterohaemorrhagic *Escherichia coli*; gbk, GenBank; IS, insertion sequence; NCBI, National Center for Biotechnology Information; PacBio, Pacific Biosciences; PHAST, PHAge Search Tool; PT, phage type; SMRT, Single Molecule Real Time; SNP, single nucleotide polymorphism; SP, Sakai prophage; Stx, Shiga toxin.

Data statement: All supporting data, code and protocols have been provided within the article or as supplementary data files.

Data Summary

The code for the pipeline can be found at: https://github.com/SharifShaaban/PROPI. All the strain sequences used to generate Fig. 1 can be found under the BioProject ID

Received 28 July 2016; Accepted 31 October 2016

PRJNA248042: https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA248042. The information related to all the other strains used in this analysis can be found in Table 1. Two supplementary tables and two supplementary figures are available with the online Supplementary Material.

Introduction

The availability of additional sequences to compare with the first sequenced *Escherichia coli* genome, *E. coli* K12 MG1655, has highlighted how the evolution of this species is intimately associated with the integration of bacteriophages into the bacterial genome, and their subsequent entrapment, recombination and degradation as prophage regions (Ohnishi *et al.*, 2001; Shaikh & Tarr, 2003). The importance of specific prophages and their longer-term legacy are evident when considering the emergence of enterohaemorrhagic *E. coli* (EHEC) as a serious zoonotic pathogen (Hayashi *et al.*, 2001; Ogura *et al.*, 2009; Perna *et al.*, 2001). EHEC are defined by their capacity to cause bloody diarrhoea and, in a subset of cases, life-threatening haemolytic uraemic syndrome (Akashi *et al.*, 1994; Tarr *et al.*, 2005). EHEC O157 : H7 is one of the main serotypes associated with disease in Europe, North America and Asia, with infections usually originating from a ruminant reservoir, particularly cattle (Naylor *et al.*, 2005), although fresh-produce outbreaks are increasingly common (Lynch *et al.*, 2009; Marder *et al.*, 2014). Phylogenetic studies have established that *E. coli* O157 can be delineated into three main lineages, as well as nine clades (Eppinger *et al.*, 2011; Zhang *et al.*, 2007; Manning *et al.*, 2008). In the USA, clade 8 strains of lineage I/II are associated with more severe human disease (Manning *et al.*, 2008); while in the UK, the main human isolates reside in lineage I and clade 4/5 (Dallman *et al.*, 2015).

Human pathology is a direct and indirect result from the activity of Shiga toxin (Stx), a two-component toxin encoded on integrated bacteriophage and released with nascent bacteriophage following cell lysis (Tyler *et al.*, 2013). In addition, EHEC are defined by the expression of a type III secretion system that injects effector proteins into epithelial cells promoting colonization of the host (Kaper *et al.*, 2004; Tobe *et al.*, 2006). Furthermore, many of these injected effector proteins are encoded by prophage regions disseminated around the genome, forming part of a prophage regulatory network that is critical for the virulence of the organism (Tree *et al.*, 2009). Prophages are integral to the evolution of EHEC O157 genomes, but relatively few studies have investigated the potential prophage variability between the main lineages and sub-clusters of this clonal serotype. Seminal work on the prophage repertoire in the EHEC O157 Sakai strain was conducted by Asadulghani and colleagues, in which they demonstrated that bacteriophage diversity can be produced from this single EHEC O157 strain following SOS-based induction (Asadulghani *et al.*, 2009). However, comparative genomics of prophage regions in EHEC isolates is hampered by the inability of short-read sequencing to resolve the large number of repetitive and paralogous features indicative of prophage sequences.

Impact Statement

Enterohaemorrhagic *Escherichia coli* (EHEC) O157 : H7 strains pose a threat to human health and are usually acquired from ruminants, the environment or fresh produce. Recent whole genome sequencing based on short-read technologies has aided outbreak tracing and has provided insights into the evolution of this pathogen. However, these methods do not capture the genomic variation that underpins differences in zoonotic and pathogenic potential. This variation is, in part, driven by the acquisition of bacteriophages (phages), which contain many similar sequences requiring longer-read sequencing technologies to define their complete composition and position in the genome. This study has used Single Molecule Real Time (SMRT) sequencing, a long-read technique, to define the integrated phage sequences in an isolate set selected to represent the wide diversity of EHEC O157. We demonstrate that the most recent diversification correlates with acquisition of phages encoding specific types of Shiga toxin, responsible for the main damage and life-threatening consequences of EHEC in humans. Smaller phage regions have preferentially lost genes that allow phage production, and the density of insertion sequence elements in integrated phage regions supports their involvement in gene deletion and phage entrapment. Profiling of integrated phages will aid identification of virulent isolates from short-read sequencing currently being adopted more routinely in diagnostic laboratories.

Table 1. Strains used in the analysis

Accession no.	Strain	BioSample ID	BioProject ID
CP018252	9000	SAMN05544760	PRJNA336330
CP018250	10671	SAMN05544761	PRJNA336330
CP018247	7784	SAMN05544762	PRJNA336330
CP018237	155	SAMN05544764	PRJNA336330
CP018243	350	SAMN05544765	PRJNA336330
CP018239	272	SAMN05544766	PRJNA336330
CP018245	472	SAMN05544767	PRJNA336330
CP018241	319	SAMN05544768	PRJNA336330
CP015832	180	SAMN05007044	PRJNA321984
NC_002695	Sakai	NA	PRJNA57781
NC_011353	EC4115	SAMN02603441	PRJNA224116
NC_013008	TW14359	SAMN02604255	PRJNA224116
CP008957	EDL933	SAMN02905113	PRJNA253471
CP010304	SS52	SAMN03265100	PRJNA201344

NA, Not applicable.

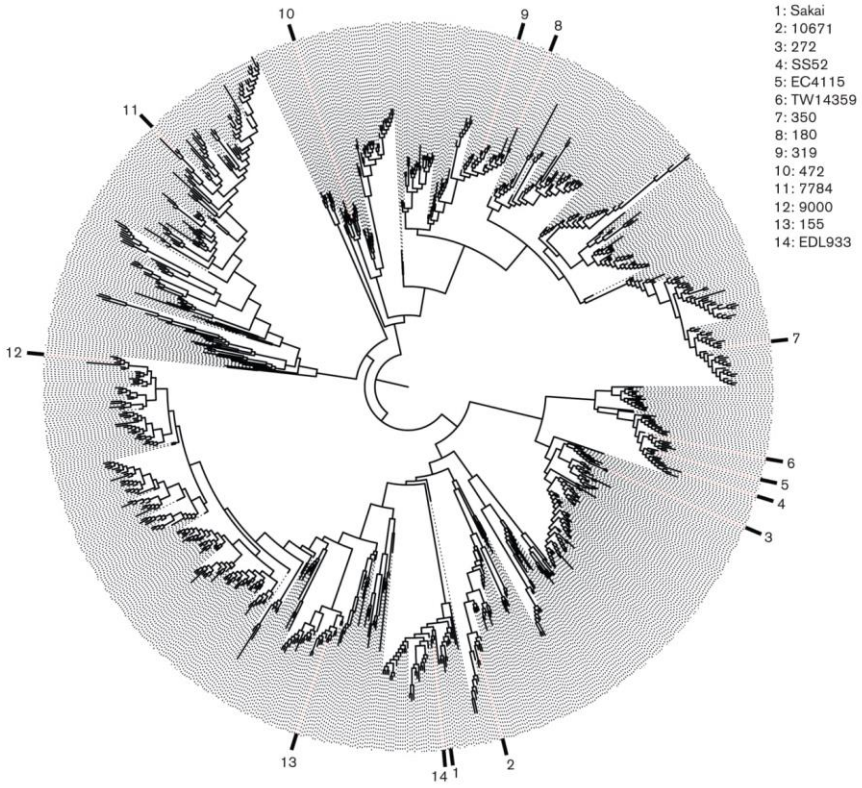


Fig. 1. Maximum likelihood phylogeny of 956 isolates representing 22 805 SNPs across 3313 coding DNA sequences (CDSs) (2569 non-coding SNPs) with a total core genome size of 3 003 626 bp. The 14 isolates analysed in the present study are labelled. Their spread across the tree demonstrates the diversity of isolates analysed, which cover multiple lineages, geographical locations and PTs.

Very recent research has provided insights into the evolution of EHEC O157 strains based on the sequencing of over 1000 isolates from human clinical cases and cattle hosts in the UK (Dallman *et al.*, 2015). This study demonstrated that the contemporary *E. coli* O157 clone emerged approximately 150 years ago from a strain harbouring a specific subtype of Stx: Stx2c. Only in the last 30–50 years was this subsequently followed by the independent acquisition of the Stx2a subtype by bacteriophage integration. Further, analysis of disease outcome indicated that more severe pathology was associated with isolates expressing Stx2a alone or in combination with Stx2c. As a consequence, it can be argued that the emergence of EHEC O157 as a serious human pathogen has coincided with the appearance of Stx2a-positive isolates in the ruminant reservoir (Dallman *et al.*, 2015).

Whilst phylogenetic analysis based on draft genomes continues to yield important insights into the epidemiology of EHEC O157, long-read sequencing platforms such as those developed by Pacific Biosciences (PacBio) and Oxford Nanopore are gaining traction, reducing the cost of complete genome sequencing and facilitating strain comparison of prophage regions (Anton *et al.*, 2015; McCarthy, 2010; Mikheyev & Tin, 2014). In the current study, we have investigated the prophage population present in 14 strains using PacBio RSII Single Molecule Real Time (SMRT) sequencing (nine isolates) and publically available genome

sequences (five strains). The strains selected to examine the prophage diversity were chosen from the different lineages and main sub-clusters previously demonstrated for *E. coli* O157 based on core genome comparison (Dallman *et al.*, 2015). Sequences for whole prophage regions were identified, extracted, clustered and compared with respect to their annotated gene content. This work has demonstrated the stability of Stx2c-encoding prophages compared to Stx2a- and Stx1a-encoding prophages, and provided general insights into the evolution of prophages in the *E. coli* genome and their interplay with insertion sequence (IS) elements.

Methods

Sequences and sequencing. For this analysis, 14 genome sequences were used. Five of these were publically available in the National Center for Biotechnology Information (NCBI) database, including genome sequences from the strains Sakai, EC4115, TW14359, EDL933 and the recently published super-shedding strain, SS52 (respective NCBI accession numbers: NC_002695, NC_011353, NC_013008, CP008957 and CP010304) (Table 1).

Sequencing of the nine isolates was conducted using a PacBio long-read sequencing RS II platform and carried out at the U. S. Department of Agriculture facility in Nebraska, USA. Qiagen Genomic-tip 100/G columns and a modified protocol, as previously described (Clawson *et al.*, 2009), were used to extract high molecular weight DNA. Using a g-TUBE (Corvaris), 10 µg DNA was sheared to a targeted size of 20 kb and concentrated using 0.45× volume of AMPure PB magnetic beads (Pacific Biosciences). Following the manufacturer's protocol, 5 µg sheared DNA and the PacBio DNA SMRTbell Template Prep kit 1.0 were used to create the sequencing libraries. A BluePippin instrument (Sage Science) with the SMRTbell 15–20 kb setting was used to size select 10 kb or larger fragments. The library was bound with polymerase P5 and sequencing was conducted with the C3 chemistry and the 120 min data collection protocol.

Assembly and annotation. SMRT analysis was used to generate a FASTQ file from the PacBio reads, which were then error-corrected using PBcR with self-correction (Koren *et al.*, 2013). The Celera Assembler was used to assemble the longest 20× coverage of the corrected reads. The resulting contigs were improved using Quiver (Chin *et al.*, 2013) and annotation was conducted using a local instance of Do-It-Yourself Annotator (DIYA) (Stewart *et al.*, 2009). Geneious (Biomatters) was used to remove duplicated sequence from the 5′ and 3′ ends to generate the circularized chromosome. Initially, OriFinder was used to determine the origin of replication (Luo *et al.*, 2014) and the chromosome was reoriented using the origin as base number one. However, for visualization purposes, the genome orientation and first base were modified to match those of the main NCBI reference sequence (Sakai) using tools from the EMBOSS tools suite (Rice *et al.*, 2000).

Phylogenetic context. To provide context for the selected strains, a phylogenetic analysis was conducted involving 943 isolates. These isolates were representatives from single linkage clusters defined at a 25 single nucleotide polymorphism (SNP) threshold from the Public Health England Shiga toxin-producing *E. coli* O157 genome collection available at the BioProject PRJNA248042 (Dallman *et al.*, 2015). As this approach requires Illumina reads, these were artificially created using wgsim from the whole genome sequences (Li *et al.*, 2009) for 13 of the 14 sequences analysed in this study (Sakai was already included). Illumina reads for all isolates (956 in total) were quality trimmed (Bolger *et al.*, 2014) and mapped to the reference EHEC O157 strain Sakai using BWA MEM (Li & Durbin, 2010). SNPs were then identified using GATK2 (McKenna *et al.*, 2010) in unified genotyper mode. Core genome positions that had a high quality SNP (>90 % consensus, minimum depth 10×, GQ ≥30) in at least one strain were extracted and RaxML v8.17 (Stamatakis, 2014) used to derive the maximum likelihood phylogeny of the isolates under the GTRCAT model of evolution (Fig. 1). This phylogeny demonstrated how the 14 selected strains for this analysis sample the diversity of EHEC O157.

Prophage calling. The PHAge Search Tool (PHAST) was used to extract prophage regions (Zhou *et al.*, 2011). Sequences were submitted using the PHAST URLAPI. Prophage regions called by PHAST, regardless of size and quality score, were extracted from their respective genomes. Any two prophages that were separated by less than 4000 bp were joined and called as a single prophage, primarily as it was observed that gaps of this size and smaller were often due to IS elements. Prior testing was carried out on PHAST's capacity to call prophage boundaries using prophages with known and studied insert sites, as well as the established Sakai prophages (SPs). The results obtained provided confidence in this approach; all SPs were found, while only two SP-like regions called as prophages. Due to the PHAST algorithm, prophage boundaries were often different to those defined for the related prophages in strain Sakai. PHAST uses an algorithm that calculates phage gene presence and distance, and the boundary is set based on short nucleotide repeats (when the phage contains an integrase), or when the distance between the phage genes is too wide (Zhou *et al.*, 2011). As a result, the extracted prophage sequences are sometimes extended beyond their physical integration sites. In addition, certain predicted prophages were found to overlap at their boundaries and these were not merged but included as separate prophages for the rest of the analyses. However, neither of these issues should have impacted on the main analyses presented.

Extracted prophages were annotated twice: once using the U. S. Department of Agriculture DIYA Glimmer-based pipeline, and the second time using Prokka (Seemann, 2014). The Prokka parameters given included a FASTA amino acid database file obtained from the previously annotated whole genomes. The former annotation method led to more hypothetical proteins and provided a gene ID usable for gene

ontology, while Prokka offered less hypothetical protein hits and more coding DNA sequence regions. Applying the DIYA annotation pipeline to the extracted prophage regions from the 14 strains resulted in a total of 9416 predicted gene products (2790 as hypothetical proteins), 523 of which were unique. This dataset had 863 unique RepIDs. These were provided as inputs for DAVID (Huang *et al.*, 2009a, b), which recognized 428 of these IDs. However, the optimal functional classification result had 347 of these as singletons (only 81 genes were grouped). By comparison, Prokka annotation yielded 13 333 gene products, of which 2106 were hypothetical proteins, with 718 unique gene products.

The prophage GenBank (gbk) files obtained from the Prokka annotation were then modified so that annotated genes were given a colour flag based on their function. Functional groups were delimited as: metabolism and transport, structure, effector and virulence factors, recombination and replication, regulation, lysis, tRNA, and hypothetical or ambiguous genes. The group into which each gene was assigned was determined using a 'key word' classification (Table S1, available in the online Supplementary Material). The selection of the key words was determined by manual curation, supplemented by the results from the DAVID analysis (data not shown). Only when a gene had a clear and studied specific function was it added to a group other than 'hypothetical and ambiguous'.

Prophage clustering. Prophages were clustered based on gene homology that was obtained using GetHomologues version 1.0 (Contreras-Moreira & Vinuesa, 2013). When running the pangenome algorithms the parameter '-t 0' was given in order to identify any gene even if it was only found in a single prophage. Paralogues were excluded using the '-e' parameter. The cluster comparison tool of GetHomologues was run on the results to obtain binary (gene presence and absence) matrices for core and pangenome. The '–T' flag was used to yield a parsimony pangenome tree (Fig. 3). This was done multiple times for different cut-offs of gene coverage and identity, jointly spanning from 70 to 95 in increments of five. The matrix applied for prophage clustering used 75 %

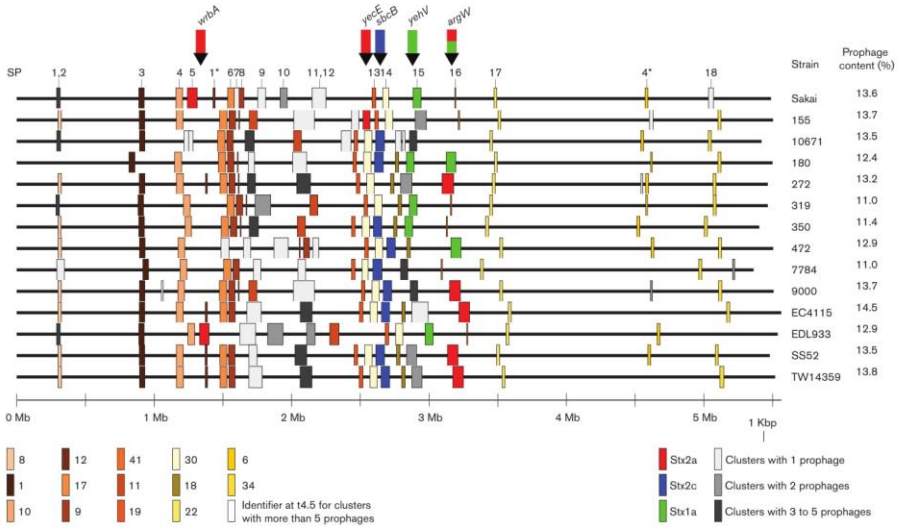


Fig. 2. Easyfig whole chromosome alignment of 14 EHEC O157 sequences with prophage regions represented as coloured blocks. The numbers and positions of the previously defined SPs are provided (Hayashi *et al.*, 2001). Blocks of the same colour show levels of similarities that approximate 80 % BLAST coverage and identity (also referred to as t4.5 Euclidean distance). At the bottom left-hand side are the cluster identifiers for each coloured block. These identifiers indicate the clusters to which each prophage belongs at t4.5 (all cluster identifiers can be linked back to individual prophages using Table S2). Therefore, blocks of the same colour indicate prophages within the same cluster, except for grey blocks which indicate that the prophage belonged to a small cluster, and Stx-encoding prophages which are coloured by Stx subtype regardless of prophage similarity. Their associated insert sites are marked with arrows above the alignment, and colour coded based on observed Stx subtypes. The coloured blocks demonstrate the large population of prophages that is conserved across strains, as well as hotspots of variation for certain prophages. The overall percentage of prophage content for each isolate was calculated and is provided on the right-hand side.

identity and coverage cut-offs. This selection was based on the default coverage cut-off of the GetHomologues tool (75 %), which allows for potential gene truncation events to be taken into account. A high identity cut-off (e.g. 95 %) is not adequate considering the multiple potential prophage families being analysed, but a low identity cut-off (e.g. 50 %) would not provide enough discrimination. Therefore, a balance of 75 % nucleotide identity was selected, which allowed, in conjunction with gene co-occurrence, for appropriate clustering of related genes (Fig. S1), but did not yield too few clusters when compared to a stricter cut-off (<20 % difference, Fig. S2).

Hierarchical clustering was performed on the resulting binary matrix. The final number of clusters selected was determined based on the maximum Euclidean distance of any two members of a cluster. Multiple maximum Euclidean distance thresholds were chosen: 0 (as a fully identical gene content threshold), 1.5, 3.0, 4.5 (as the main comparative distance) and 6.0. The primary analysis distance of 4.5 was chosen as it translates to ~80 % BLAST sequence coverage and similarity. The other thresholds were used to establish the different levels of relationship between the clusters. With these thresholds, each prophage was assigned a code consisting of five values. These values indicated in which cluster the prophage was positioned at each threshold. These codes then served as a convenient and qualitative measure of prophage relationships (Table S2).

Whole genome and prophage comparison. Whole genome alignments were conducted with Easyfig (Sullivan *et al.*, 2011). The gbk files were modified so that prophages were represented as blocks with different colours based on the prophage code described above (Fig. 2). Easyfig alignments were also conducted on selected groups of prophages based on their clusters, inter-cluster relationships and Stx subtypes (Figs 4 and S1).

Stx and IS calling. The FASTA sequences of all the genomes analysed were concatenated in a multi-FASTA file and made into a BLAST database using BLAST+ version 2.2.29 (Altschul *et al.*, 1990; Camacho *et al.*, 2009). The Stx reference sequences obtained from Scheutz *et al.* (2012) were used as the query in a BLASTN comparison, with only the best scoring hits kept per Stx subtype.

ISs were identified using a BLAST database, extracted from IS Finder (Siguier *et al.*, 2006) (http://www-is.biotoul.fr),

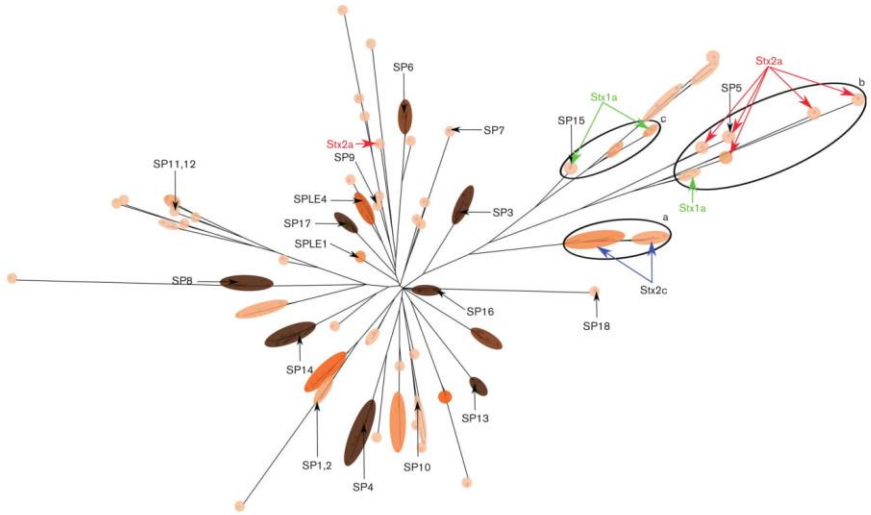


Fig. 3. Midpoint rooted parsimony tree based on the gene content of 232 prophage sequences. SP numbers and Stx-encoding prophages are marked on the tree. The coloured overlay represents the clustering at t4.5 (approximately 80 % BLAST coverage and identity). The shade and colour of the overlay represents the number of prophages within these clusters, with darker ovals representing more isolates within a cluster. The size of the oval indicates how diverse members of a cluster are, with larger ovals being more diverse. This graphic indicates that the clusters that are present in the majority of the isolates analysed (dark brown, SPs 3, 4, 6, 7, 8, 13, 14 and 16) are often closely associated 'tight clusters' with little variation. Circles (labelled a–c) determine three groups of clusters that appear to relate to Stx-encoding prophages, with Stx2a prophages exhibiting significant diversity.

containing IS elements that originated from *E. coli* and *Shigella* (total of 119 sequences) (IS Finder accessed October 2015). IS regions were determined within prophages using the following BLAST parameters: '-evalue 1e-100 -best_hit_score_edge 0.0001 -best_hit_overhang 0.25 -outfmt 6'. In order to avoid repeated IS hits from closely related IS sequences, this was followed by the filtering of hits with a minimum match length of over 700 bps, and hits with the same starting or ending position were collapsed to the one with the highest bit score. These IS regions were given a colour flag in the prophage gbk files. However, it should be noted that the single IS elements highlighted in this study usually contained two to three individual genes.

Gene frequencies and graphs. Based on the annotated functional groupings described in the 'Prophage calling' section of Methods, gene content against the mean length of prophages was plotted in R with ggplot2 (Wickham, 2009). The mean number of genes for each annotated function within each cluster at a Euclidean distance of 4.5 was calculated. A multiplication step was then applied to this so that values could be computed correctly and easily in Linux. For this, the mean gene number was multiplied by the maximum prophage length in bp. The pseudo-proportions were then determined by dividing the result by the mean prophage length of the specific cluster (bp). These were then plotted in R as a scatter plot and a Loess line drawn to visualize the line of best fit. To determine the significance of these trends a Rho Spearman rank test was conducted in R (Fig. 5).

Selecting spontaneous-cured lysogens using a temperature sensitive plasmid with TcR under CI control. The tetracycline resistance gene with native Ribosome Binding Site (RBS), but without a promoter, was amplified from pBR322 using primers Nt-pTOF24-TcR (5′-aaactgcagagatcttaacgcagtcaggcaccgtgtatg-3′) and Ct-pTOF24-TcR (5′-aaactcgagcgaggtgccgccggcttccattca-3′) and cloned into pTOF24 (Merlin *et al.*, 2002), following restriction with BglII and XhoI. Screening was carried out on chloramphenicol resistant (CmR) colonies with the same primers; the bacteria were still sensitive to tetracycline at this point because no promoter was cloned. The oL/pL promoters from the relevant Stx-encoding prophage were then amplified with primers (as below) from a lysogen and cloned into pTOF–TcR with 5′-PstI and 3′-BglII after an In-Fusion kit (Clontech) cloning step for the amplicons. Selection was for CmR. The primers for the Stx2c-prophage were 5′-Sp5pL-PstI-IF (5′-gtctcggtacccgacctgcagcctctcgcccaaaaaaacacataac-3′) and 3′-Sp5pL-BglII-IF (5′-tgcctgactgcgttaagatcttgccagtctgttccatttggcttcc-3′). The primers used for the Stx2a-prophage were 5′-560stx2pL-PstI-IF (5′-gtctcggtacccgacctgcagctttgcctcacgttcgccacc-3′) and 3′-560stx2pL-BglII-IF (5′-tgcctgactgcgttaagatcttcctgctgacgatgataataatg-3′) with the constructs used on isolate 9000. A check was then carried out to determine whether the level of TcR was lower in a lysogen compared to a non-lysogen background. When spontaneous excision of the phage occurs, the isolate reverts to a higher level of TcR as pL is activated due to a lack of CI repression. Lysogen curing was verified by PCRs across insert junctions. The final step was to remove pTOF24-oL/pL-TcR from the spontaneous lysogen-cured isolate based on growth under restrictive temperature conditions, without antibiotic.

Results

Prophage content and location in *E. coli* O157 genomes

The strains selected for prophage analysis in this study are highlighted in Fig. 1 and sampled the wide diversity of the phylogeny, including lineages and susceptibility to a panel of typing phages that define phage type (PT). We detected 232 prophages and extracted their sequences. The positions and sizes of the prophages in their respective genomes are illustrated in Fig. 2, which indicates the more stable prophage populations with consistent locations across the genomes. Prophage size ranged from 6126 to 152 606 bp and the estimated prophage content per genome ranged from 11.0 to 14.5 % (Fig. 2). Many of the prophages matched those originally designated as SPs (Hayashi *et al.*, 2001). Prophage distribution across the chromosome was biased to regions between bp 800 000 and 3 600 000, with ~83 % of all prophages found within these boundaries. All Stx-encoding prophages were found in previously documented insert sites (Shaikh & Tarr, 2003): prophages encoding Stx2c were found only in *sbcB*; while those encoding Stx2a were found in multiple insert sites – *wrbA*, *argW*, *yecE*; and Stx1a-encoding prophages were detected in *yehV* and *argW*. The majority of strains contained only one or two Stx-encoding prophages.

Prophage clustering

The extracted prophages were clustered according to related gene content. There was no single gene common to all the analysed prophages at a cut-off of 75 % coverage and nucleotide identity; this reflects the wide diversity of temperate phage backgrounds integrated into the *E. coli* O157 genome. Using the outputted binary matrix and with selected Euclidean distance thresholds of 6.0, 4.5, 3.0, 1.5 and 0, hierarchical clustering yielded 42, 63, 86, 128 and 151 clusters, respectively (full clustering in Table S1). This indicated that within our 232 prophages there were 151 different individual prophages based on gene content alone. Based on subsequent alignments (Figs 4 and S1), the majority of these clusters appeared adequate, although difficulties could arise when analysing the smaller prophages (6000–9000 bp) in this dataset, because their relatively low gene content results in less discriminatory power. The main benefit of this approach was that the relatedness of different prophages and their clusters could be visualized at different relative scales.

The midpoint rooted parsimony-based pangenomic tree obtained from GetHomologues is shown in Fig. 3, with

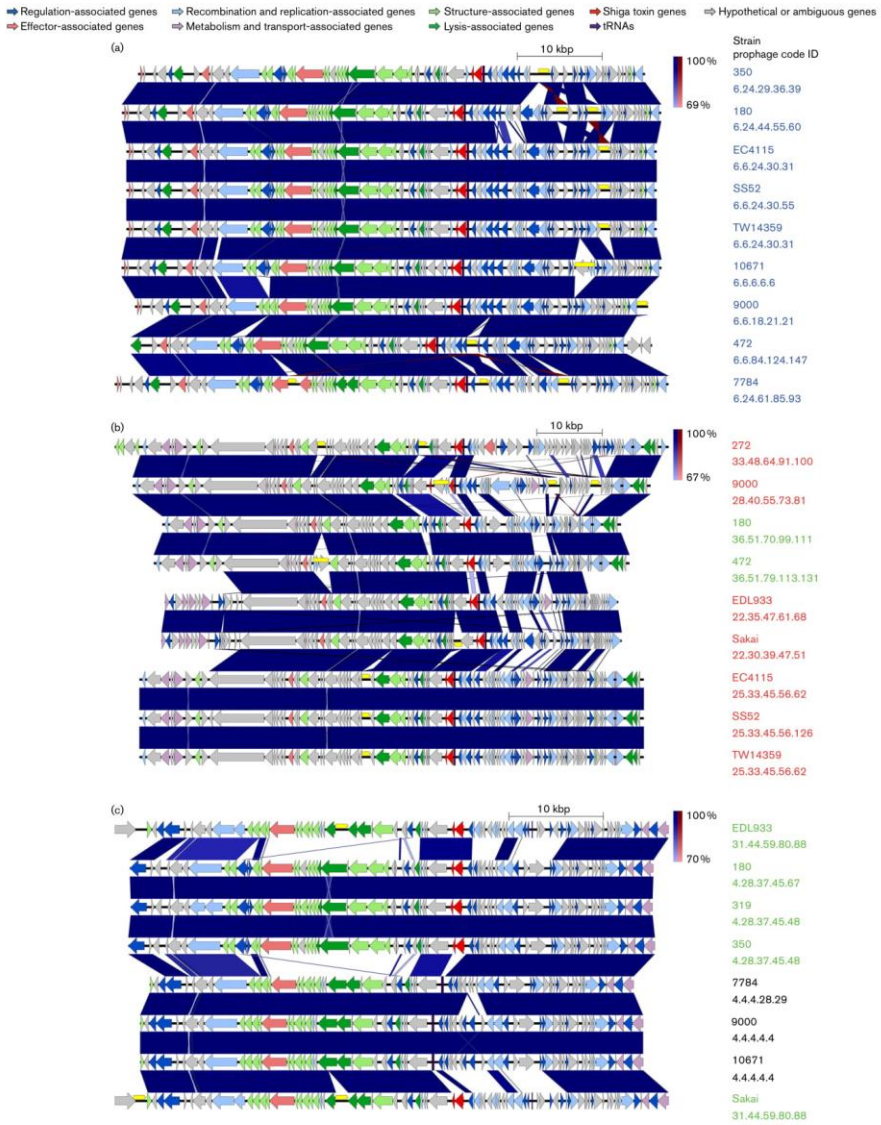


Fig. 4. Easyfig alignment of the prophages contained in the circles (a–c) in Fig. 3. Genes are colour coded based on functional groups as indicated at the top of the figure. Prophage names are colour coded based on Stx content (blue, Stx2c; red, Stx2a; green, Stx1a; black, no Stx). The prophage code shown is based on Euclidean distance thresholds of 6.0, 4.5, 3.0, 1.5 and 0, and also indicates the relationship

between the aligned prophages based on gene content. The alignments confirm the relationships shown in Fig. 3, with Stx2c-encoding prophages exhibiting a high degree of conservation, while Stx2a and Stx1a exhibit multiple subpopulations. IS elements are shown in yellow above the genes and are seen to interrupt multiple genes across the different prophages, including Stx2a in strain 9000 (centre alignment).

clusters shaded within a Euclidean distance threshold of 4.5 (t4.5). Multiple cluster types were observed including singletons, clusters with only a single prophage (SP18 and SP7). Typically, these were either at the start or at the end of a branch, indicating that they either lacked the genes defining the other cluster(s) on that branch if at the start of the branch, or possessed additional gene content to the other branch members if at the end of the branch. There were also specific singletons that appeared to be prophages integrated within other prophages. There were prophage clusters with representatives in nearly all the strains analysed (indicated by the darker shading in Fig. 3) and these were often tightly grouped, for example SP13, SP16 and SP17. Other clusters had intermediate representation in the analysed strains and these could be more divergent (SP10 and SP11, 12).

Stx-encoding prophage comparison

All but one (155 Stx2a) of the Stx-encoding prophages were present on the same branch off the main root of the prophage tree (Fig. 3). These diverged into prophage sub-clusters that associated well with specific Stx subtypes. It was noted that not all prophages within these branches necessarily carried Stx genes. Therefore, Stx-negative prophages could be found with very similar gene content to Stx-encoding prophages. Stx2a-encoding prophages were the most divergent, containing five sub-clusters (Fig. 3, circle b) based on only seven prophages. By comparison, Stx1a-encoding prophages were also relatively divergent with three sub-clusters from seven prophages (Fig. 3, circles b and c). Finally, Stx2c-encoding prophages exhibited the least diversity, with only two sub-clusters based on nine prophages (Fig. 3, circle a).

The main prophage clusters associated with *stx* were further compared by Easyfig sequence alignment (Fig. 4). The least sequence and gene content diversity was shown by the Stx2c-encoding prophages (Fig. 4a), where the main variation was associated with IS element insertion (designated by yellow blocks in Fig. 4). The prophages were assigned a code based on the clustering at the different Euclidean distances (Fig. 4). The Stx2c-encoding prophages were all within the same t6.0 cluster and two t4.5 clusters. By contrast the Stx2a-encoding prophages were more variable and the prophage groupings also contained two Stx1a-encoding prophages (Fig. 4b). There were four t6.0 and five t4.5 in relation to *stx2a* alone. The American reference genomes EC4115, SS52 and TW14359 appeared to have near identical Stx2a-encoding prophages (identical up to t1.5), while the Sakai and EDL933 Stx2a-encoding prophages were related to each other up to t6.0. The Stx2a-encoding prophages demonstrated high levels of variation in the final third of the prophage, often starting adjacent to the *stx* locus. The demonstration of two Stx1a-encoding prophages in these sub-clusters is indicative of acquisition of a different Stx subtype by a similar prophage background. This aligns with their insertion in *argW*, which occurs for a subset of the Stx2a-encoding prophages, and shows that the prophage background is more relevant to the insert site than *stx* content.

The remaining Stx1a-prophages clustered separately, consisting of two t6.0 and two t4.5 clusters (Fig. 4c). The first subtype was present in EDL933 and Sakai, while the second was in UK isolates 180, 319 and 350. A related third subtype was Stx negative and more closely related to the sub-cluster present in EDL933 and Sakai.

Table 2. The PT of the original strain 9000 and, when modified, a description of the modifications

Strain	Description	PT
9000	Original PT21/28 IPRAVE isolate, Stx2a and Stx2c	21/28
9000–2	9000 with Stx2c phage partly deleted	21/28
9000–3	9000 with Stx2a phage entirely deleted	32
9000–4	9000–2 with Stx2a phage entirely deleted	32
9000–5	9000–3 with Stx2c phage partly deleted	32

Integration of a Stx2a-encoding prophage can confer a switch from PT32 to P21/28

A common PT, PT21/28, associated with human infection in the UK, usually contains both Stx2a- and Stx2c-encoding prophages (Dallman *et al.*, 2015). By contrast, a phylogenetically close sub-cluster of PT32 contains predominantly Stx2c-encoding prophages only (Matthews *et al.*, 2013; Dallman *et al.*, 2015). We therefore tested whether excision of the Stx-encoding prophages from PT21/28 isolate 9000 (Fig. 1, Table 2) could alter the PT, which is based on resistance or susceptibility to a characterized set of T4 and T7 lytic phages (Cowley *et al.*, 2015). To generate these strains, we took advantage of the fact that lambda-like prophages continually express CI to remain in the lysogenic state. We reasoned that if we placed the prophage specific CI-repressed promotor in front of an antibiotic-resistance gene (in this case tetracycline resistance) on a plasmid transformed into isolate 9000 (PT21/28), then we would only get tetracycline resistance if the relevant prophage excised, relieving CI-based repression of the tetracycline resistance. This system worked well for the P_R promoters and we were able to routinely obtain spontaneous excision of the Stx2a-encoding prophage from isolate 9000. By contrast, a 9000 variant with a completely excised Stx2c-prophage was not obtained; instead partial deletions occurred that were

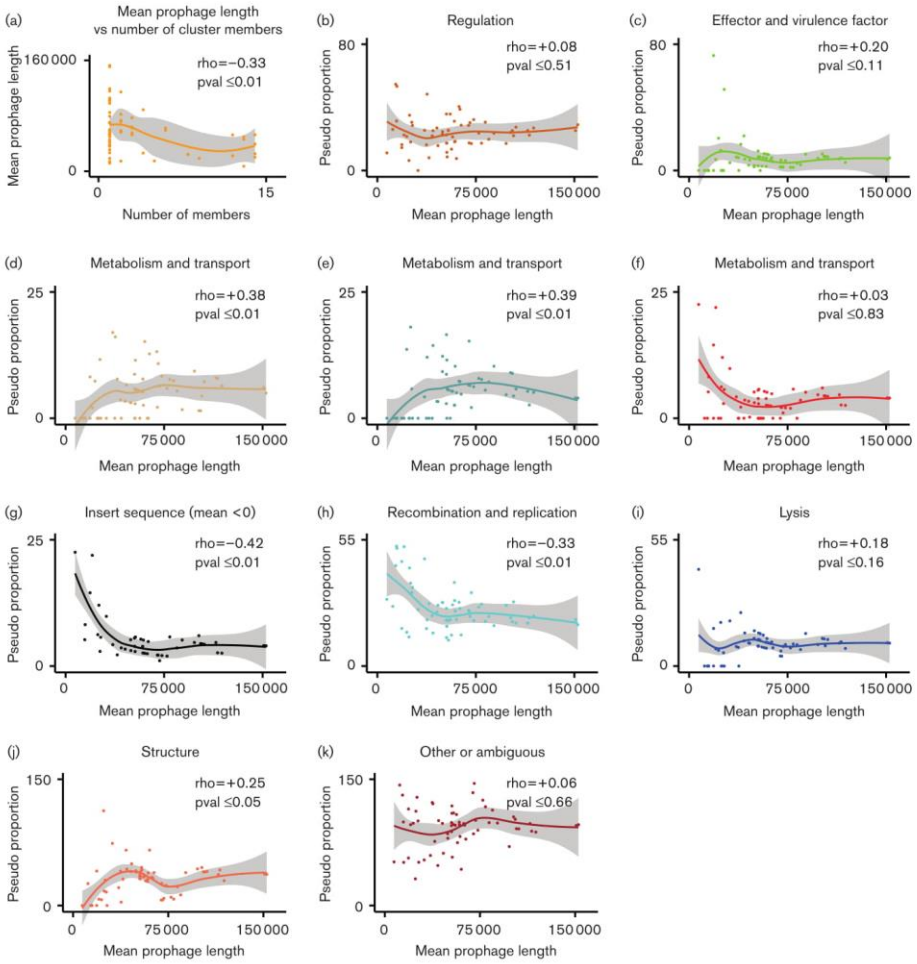


Fig. 5. (a) The mean prophage length of a cluster over its number of members, indicating a statistically significant association with more members for shorter prophage sequences ($P \leq 0.01$). (b–k) All plots represent the proportion of genes from a functional group in a cluster over its mean prophage length. (b, c, f, i and k) These show no statistical significance between their gene function groups and mean prophage length. (d, e, g, h and j) These show statistical significance relating these particular function groups to the mean prophage length of clusters ($P \leq 0.05$).

missing the main phage lysis/lysogeny regulatory proteins and the *stx*2c genes (data not shown). A double deletion was attempted twice, once by first selecting for Stx2a-prophage excision and then Stx2c-prophage excision (partial), and the second time in the reverse order. Phage typing of the resultant strains demonstrated that any strain that had excised the Stx2a-prophage had a PT32 designation, rather than the PT21/28 of the parent strain (Table 2).

Table 3. The number of gene products falling in each of the functional categories established in Methods

The 'All gene products' column enumerates all combined gene products including repeats, while the 'Unique gene products' column does not include duplicates.

Functional group	All gene products	Unique gene products
Regulation	1500	136
Effector and virulence factors	564	45
Metabolism and transport	302	44
tRNA	429	429
Recombination and replication	1629	109
Lysis	686	28
Structure	2174	108
Hypothetical or ambiguous	6001	228

IS elements

IS activity is critical to the evolution of individual isolates. For example, isolate 9000 had an ISec8 (IS*66* family) within the *stx*2a subunit A gene that is likely to disrupt Stx2a production (Fig. 4b). The same isolate also had an IS*629* (IS*3* family) in a flanking excisionase of the Stx2c-encoding prophage that could impact on prophage induction (Fig. 4a). Other IS activity can be seen in Figs 4 (and S1) clearly delineated by gaps in the BLAST alignment. IS elements appear in regions of recombination and change, with a bias towards prophage regions (Ooka *et al.*, 2009). We found 277 IS elements in the prophage regions of our strains, with 401 IS elements across the whole chromosome regions; the latter based on the same method of extraction as for prophage regions but using whole genome sequences (this can, however, include highly similar IS BLAST hits in the same location). This translates to 69.1 % of IS elements in 12.9 % of the whole genomes (prophage regions).

Statistical analysis of prophage content and size

We hypothesized that as these *E. coli* genomes contained prophages that were acquired at different times, older prophages will have been under the evolutionary pressure of the host isolate for longer and should be more host adapted. Certain SPs are cryptic, i.e. can no longer produce viable bacteriophages (Asadulghani *et al.*, 2009), but presumably this process from productive prophage to cryptic prophage is a continuum, in which genes useful to the host bacterium may be retained and those specific to bacteriophage production are lost. We investigated this in two ways; in the first we asked whether there was a correlation between the mean length (bp) of a prophage cluster (defined at t4.5) and frequency in the analysed strains. As indicated in Fig. 5(a), shorter prophages are more likely to be associated with a greater number of strains ($P \leq 0.01$). The second approach examined the correlation between functional gene content and the mean length (bp) of prophage clusters at t4.5 (Fig. 5b–k).

From the analysis, the prophage sequences yielded 13 333 potential proteins, of which 2106 were hypothetical. These were sorted into 718 unique gene products that were used for functional groupings (Table 3), and the key words used for functional assignments can be found in Table S1. Genes functionally annotated as regulatory (Fig. 5b), effectors and virulence factors (Fig. 5c), lytic (Fig. 5i) and 'other' (Fig. 5k) showed no significant correlation with prophage length. IS element content was then examined in all prophages and no correlation was evident (Fig. 5f). However, we noted that IS elements were completely absent from a subset of prophage clusters, so we examined whether the proportion of IS element content correlated with prophage length for clusters infected with at least one IS element (Fig. 5g). In this case there was a correlation, with shorter prophages containing a relatively higher proportion of IS elements. A similar correlation was seen for recombination and replication-associated genes in all prophages (Fig. 5h). The proportions of three functional groupings showed evidence of a reduction as the mean prophage length shortened. These were metabolism and transport (Fig. 5d), tRNAs (Fig. 5e) and structural, predominately phage tail, head and baseplate, genes (Fig. 5j).

Discussion

This study aimed to analyse prophage regions extracted from *E. coli* O157 : H7 strains representative of the main clusters found in the extensive phylogeny presented by Dallman *et al.* (2015). This phylogeny was based on core sequence analysis, but does not provide an indication of how the accessory genome, in particular prophage composition, may vary. Central to EHEC O157 virulence in humans is the subtype of Stx and control of Stx production, intrinsically linked to the prophage in which it is encoded. The original model of Stx acquisition hypothesized that Stx1-encoding prophages were the first to be acquired by EHEC O157, followed by Stx2-encoding prophages. However, due to the presence of differing Stx subtypes, the situation is more complex: a recent study looking at the phylogeny of over 1000 EHEC O157 isolate was suggestive that the original acquisition of Stx2c-encoding prophages by EHEC O157 occurred about 150–175 years ago, followed by later acquisition of Stx2a and Stx1a (Dallman *et al.*, 2015). However, this analysis was limited by the drawbacks of short-read sequencing. In our current study, we have been able to extract and examine a large number of prophage sequences from 14 strains, including those encoding Stx. Stx2c-encoding prophages were found to be less variable in their gene content and sequence similarity across strains from different geographical locations and lineages, compared to more diverse Stx1a- and Stx2a-encoding prophages.

The core genome phylogeny study (Dallman *et al.*, 2015) indicated that PT21/28, a PT associated with the majority of serious human EHEC infections over the last decade, emerged from a PT32 progenitor. In the present study, we demonstrated that excision of a Stx2a-encoding prophage from a PT21/28 isolate (9000) results in the isolate being defined as PT32. This is due to resistance to typing phages 6 and 13, which is a facet of the PT21/28 phenotype, but not the PT32 phenotype (Cowley *et al.*, 2015). It should be noted that clades of PT32 exist containing both Stx2a- and Stx2c-encoding prophages (Dallman *et al.*, 2015), indicating that conversion with a Stx2a-encoding prophage does not necessarily always lead to an altered lytic phage susceptibility. This presumably reflects the specific cluster of Stx2a-prophage involved in the lysogenization with a subset causing this typing transition. Our clustering of Stx2a-encoding prophages shows these to be one of the most diverse prophages present in the *E. coli* O157 clonal complex, with variation in the regulatory regions underpinning the different Stx2a-encoding prophage clusters. Numerous reports in the literature have observed these multiple Stx2a-encoding prophage subtypes (Ogura *et al.*, 2009; Yin *et al.*, 2015), with one paper classifying these subtypes using PCRs across the prophage regulatory region (Ogura *et al.*, 2015). The Stx2a-encoding prophages presented in our analyses can generally be classified into their groupings (alpha to zeta) by *in silico* approaches using primer recognition (data not shown). However, the PCR method places Stx2a-encoding prophages of strains EC4115, SS52, TW14359 and EDL933 in the same group, while Euclidean clustering groups the ones of EDL933 closer to Sakai, because it also takes into account differences outside the regulatory region. In addition, the Stx2a-encoding prophage of isolate 155 did not classify into any of their predefined groups, including 'untypable'; thus, clearly demonstrating the marked difference in that particular Stx2a-encoding prophage compared to others so far studied.

Our methods also found two Stx1a-encoding prophages in isolates 180 and 472, which clustered closely to those usually encoding Stx2a. Both of these were defined as gamma type (Ogura *et al.*, 2015) (in one case both primers were found, while in another only one was found), suggesting that different Stx subtypes have been acquired by closely related temperate bacteriophage types, although whether this can impact on the pathogenic potential of the isolate is unknown. Many more Stx2a- and other Stx-encoding prophage sequences are required, along with metadata around the associated infections, in order to fully analyse the implications of the different Stx-encoding prophage subtypes. A longer-term objective of our work is to develop methods to define prophage content from short-read sequence data in order to help predict pathogenic potential. However, this will demand a deeper understanding of the current prophage populations generated by long-read sequencing methods, and more access to disease-severity data as well as to other epidemiological traits associated with the sequenced isolates.

A seminal research paper for EHEC was published in 2009 (Asadulghani *et al.*, 2009), examining the Sakai strain prophage content and capacity of such regions to be circularized and form infective phage particles. It was observed that many of the prophage regions contained mutations and deletions that should inhibit excision/replication and/or phage production. Of the 18 SPs, 9 were shown to be excisable either spontaneously or following mitomycin induction, whereas 6 SPs were packaged in a DNAse-resistant manner and 4 prophage regions could be transduced into *E. coli* K12. Although this indicated that several prophages considered cryptic could be excised, it was also apparent that nine showed no evidence of excision. From our study three of these 'fixed' prophages were present in most of our analysed strains at a high level of similarity (SP3, SP8 and SP17), three others (SP1, SP2 and SP16) were limited in diversity and only found in subsets of our strains, while the remaining three (SP11, SP12 and SP18) were only found in Sakai. Conversely, we have shown that the most variation in prophage regions was present in excisable prophages, which may be due to increased recombination activity during replication. Specifically, in the Asadulghani *et al.* (2009) study, generation of a hybrid Stx1-encoding prophage was identified, based on the SP5 background, indicating exchange of the Stx1-encoding region into the SP5 prophage. As previously stated, we noted that in our study, two of the UK isolates (180 and 472) contained SP5-like prophages, but encoded Stx1a (Fig. 4b), supporting the potential occurrence of this recombination in wild-type isolates. It was also apparent that the majority of variation present in Stx2a-encoding prophages (SP5-like) occurred in a region that encodes Stx2a regulatory and accessory genes, but distinct from structural and lytic genes (Fig. 4b).

We propose a model of prophage entrapment and 'fixing' based on the activity of IS elements in the *E. coli* O157 genome. This would then be followed by attrition/loss of specific genes that are no longer of value to the bacterium or phage, because it subsequently cannot produce viable bacteriophage (at least without helper phage activity). In support of IS 'entrapment', we observed IS insertion into the excisionase (*xis*) of the Stx2c-encoding prophage of isolate 9000, which would be expected to prevent prophage excision from the genome. It was also evident from our work that the Stx2c-prophage does not excise cleanly from isolate 9000, although the molecular basis for this requires elucidation. In the Asadulghani *et al.* (2009) study, they also noted that several prophages (including SP4 and SP14) contained truncated excisionases. It is proposed that combinatorial IS activity will result in rearrangements and deletions: we observed that prophage regions in our strains were biased for IS insertions, as has been noted previously (Ooka *et al.*, 2009). Alignment of the majority of the prophage regions (Figs 4 and S1) indicated that much of the variation in conserved prophages was due to IS integration.

In our study, we carried out functional gene annotation of the prophage regions to assess whether particular functional groups were retained or lost as prophages co-evolved with

the bacterium. Specifically, we hypothesized that shorter prophages may represent elements that have been trapped for longer in the genome and may have undergone more gene deletion events. In support of this, there was a negative correlation between prophage length (bp) and their frequency in the strains, indicating that shorter prophages were more commonly found in the majority of strains. In addition, certain functional groups were preferentially lost in relation to prophage size, including structural genes that are predominately phage associated, such as head, tail and baseplate genes. This agrees with the concept that these genes may no longer be of value once the prophage is 'fixed'. Lysis genes showed no significant trend, which may reflect a requirement for their maintenance or simply a need for more sequences, as the number of such genes was limited. Effectors and virulence factors showed no significant trend with size. Functional groups that were preferentially maintained included recombination and replication, which in conjunction with the trend observed for maintenance and spread of IS elements is indicative of the value of these genes in the bacterium. IS elements and recombinases generate diversity that may be a critical factor for their retention. The strains we analysed were from different geographical locations and collected at different times, yet contained many prophages that were nearly identical between all the strains. Therefore, while we are viewing only evolutionary snapshots, it would appear that this evolution follows a model where genes get either entrapped or lost from an isolate, and this then becomes a key representative isolate within a population.

A further observation from our study was the presence of an IS element within the Stx2a A subunit of isolate 9000. A number of studies have demonstrated similar integration and excision into and out of *stx* genes (Asano *et al.*, 2013; Park *et al.*, 2013; Toro *et al.*, 2015). Under certain conditions the loss of Stx prophages and Stx activity has been demonstrated, indicating that Stx activity may be a negative factor in certain cases and an advantage in others (Park *et al.*, 2013). IS elements, therefore, have the potential to provide heterogeneity within populations, further driving their maintenance in, or close to, strongly selective loci. Further work is required to examine IS element stability, and the effects on gene expression.

In summary, the identified prophage population is diverse but can be classified, potentially allowing clusters to be called from short-read sequencing data. By continuing to expand our prophage database from long-read approaches, we aim to be able to provide prophage profiles (similar to Fig. 2) that can have predictive value when coupled with epidemiological metadata. In this way, we aim to extract both 'core' SNP data and accessory prophage data from reads for diagnostic and public-health benefit.

Acknowledgements

This work was funded by a Food Standards Agency/Food Standards Scotland research programme (FS101055), and supported by a Biotechnology and Biological Sciences Research Council strategic programme (BB/J004227/1) at the Roslin Institute. We would like to acknowledge the value of bovine *E. coli* isolates collected as part of a previous Wellcome Trust funded IPRAVE programme, as well as isolates and advice from the Scottish *E. coli* Reference Laboratory (SERL). We would also like to thank Dr G. Devailly (Roslin Institute) for his guidance and help with the R scripts.

References

Akashi, S., Joh, K., Tsuji, A., Ito, H., Hoshi, H., Hayakawa, T., Ihara, J., Abe, T., Hatori, M. & other authors (1994). A severe outbreak of haemorrhagic colitis and haemolytic uraemic syndrome associated with *Escherichia coli* 0157 H7 in Japan. *Eur J Pediatr* **153**, 650–655.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403–410.

Anton, B., Mongodin, E., Agrawal, S., Fomenkov, A., Byrd, D., Roberts, R. & Raleigh, E. (2015). Complete genome sequence of ER2796, a DNA methyltransferase-deficient strain of *Escherichia coli* K-12. *PLoS One* **10**, e0127446.

Asadulghani, M., Ogura, Y., Ooka, T., Itoh, T., Sawaguchi, A., Iguchi, A., Nakayama, K. & Hayashi, T. (2009). The defective prophage pool of *Escherichia* coli O157: prophage-prophage interactions potentiate horizontal transfer of virulence determinants. *PLoS Pathog* **5**, e1000408.

Asano, Y., Karasudani, T., Tanaka, H., Matsumoto, J., Okada, M., Nakamura, K., Kondo, H. & Shinomiya, H. (2013). Characterization of the *Escherichia coli* O157:H7 outbreak strain whose Shiga toxin 2 gene is inactivated by is 1203v insertion. *Jpn J Infec Dis* **66**, 201–206.

Bolger, A. M., Lohse, M. & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120.

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

Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J. & other authors (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**, 563–569.

Clawson, M. L., Keen, J. E., Smith, T. P., Durso, L. M., McDaneld, T. G., Mandrell, R. E., Davis, M. A. & Bono, J. L. (2009). Phylogenetic classification of *Escherichia coli* O157:H7 strains of human and bovine origin using a novel set of nucleotide polymorphisms. *Genome Biol* **10**, R56.

Contreras-Moreira, B. & Vinuesa, P. (2013). GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl Environ Microbiol* **79**, 7696–7701.

Cowley, L. A., Beckett, S. J., Chase-Topping, M., Perry, N., Dallman, T. J., Gally, D. L. & Jenkins, C. (2015). Analysis of whole genome sequencing for the *Escherichia coli* O157:H7 typing phages. *BMC Genomics* **16**, 271.

Dallman, T. J., Allison, L., Gally, D. L., Wain, J., Ashton, P. M., Petrovska, L., Woolhouse, M. E. J., Jenkins, C., Byrne, L. & other authors (2015). Applying phylogenomics to understand the emergence of Shiga-toxin-producing *Escherichia coli* O157:H7 strains causing severe human disease in the UK. *Microb Genom* **1**.

Eppinger, M., Mammel, M. K., Leclerc, J. E., Ravel, J. & Cebula, T. A. (2011). Genomic anatomy of *Escherichia coli* O157:H7 outbreaks. *Proc Natl Acad Sci U S A* **108**, 20142–20147.

Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C. G., Ohtsubo, E., Nakayama, K. & other authors (2001). Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res* **8**, 11–22.

Huang, D. W., Sherman, B. T. & Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1–13.

Huang, D. W., Sherman, B. T. & Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44–57.

Kaper, J., Nataro, J. & Mobley, H. (2004). Pathogenic *Escherichia coli*. *Nat Rev Microbiol* **2**, 123–140.

Koren, S., Harhay, G. P., Smith, T. P., Bono, J. L., Harhay, D. M., Mcvey, S. D., Radune, D., Bergman, N. H. & Phillippy, A. M. (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol* **14**, R101.

Li, H. & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079.

Luo, H., Zhang, C. T. & Gao, F. (2014). Ori-Finder 2, an integrated tool to predict replication origins in the archaeal genomes. *Front Microbiol* **5**, 482.

Lynch, M. F., Tauxe, R. V. & Hedberg, C. W. (2009). The growing burden of foodborne outbreaks due to contaminated fresh produce: risks and opportunities. *Epidemiol Infect* **137**, 307–315.

Manning, S. D., Motiwala, A. S., Springman, A. C., Qi, W., Lacher, D. W., Ouellette, L. M., Mladonicky, J. M., Somsel, P., Rudrik, J. T. & other authors (2008). Variation in virulence among clades of *Escherichia coli* O157:H7 associated with disease outbreaks. *Proc Natl Acad Sci U S A* **105**, 4868–4873.

Marder, E. P., Garman, K. N., Ingram, L. A. & Dunn, J. R. (2014). Multi-state outbreak of *Escherichia coli* O157:H7 associated with bagged salad. *Foodborne Pathog Dis* **11**, 593–595.

Matthews, L., Reeve, R., Gally, D. L., Low, J. C., Woolhouse, M. E., McAteer, S. P., Locking, M. E., Chase-Topping, M. E., Haydon, D. T. & other authors (2013). Predicting the public health benefit of vaccinating cattle against *Escherichia coli* O157. *Proc Natl Acad Sci U S A* **110**, 16265–16270.

McCarthy, A. (2010). Third generation DNA sequencing: Pacific Biosciences' single molecule real time technology. *Chem Biol* **17**, 675–676.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S. & other authors (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303.

Merlin, C., McAteer, S. & Masters, M. (2002). Tools for characterization of *Escherichia* coli genes of unknown function. *J Bacteriol* **184**, 4573–4581.

Mikheyev, A. S. & Tin, M. M. (2014). A first look at the Oxford Nanopore MinION sequencer. *Mol Ecol Resour* **14**, 1097–1102.

Naylor, S. W., Roe, A. J., Nart, P., Spears, K., Smith, D. G. E., Low, J. C. & Gally, D. L. (2005). *Escherichia coli* O157: H7 forms attaching and effacing lesions at the terminal rectum of cattle and colonization requires the *LEE4* operon. *Microbiology* **151**, 2773–2781.

Ogura, Y., Ooka, T., Iguchi, A., Toh, H., Asadulghani, M., Oshima, K., Kodama, T., Abe, H., Nakayama, K. & other authors (2009). Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic *Escherichia coli*. *Proc Natl Acad Sci* **106**, 17939–17944.

Ogura, Y., Mondal, S. I., Islam, M. R., Mako, T., Arisawa, K., Katsura, K., Ooka, T., Gotoh, Y., Murase, K. & other authors (2015). The Shiga toxin 2 production level in enterohemorrhagic *Escherichia coli* O157:H7 is correlated with the subtypes of toxin-encoding phage. *Sci Rep* **5**, 16663.

Ohnishi, M., Kurokawa, K. & Hayashi, T. (2001). Diversification of *Escherichia coli* genomes: are bacteriophages the major contributors? *Trends Microbiol* **9**, 481–485.

Ooka, T., Ogura, Y., Asadulghani, M., Ohnishi, M., Nakayama, K., Terajima, J., Watanabe, H. & Hayashi, T. (2009). Inference of the impact of insertion sequence (IS) elements on bacterial genome diversification through analysis of small-size structural polymorphisms in *Escherichia coli* O157 genomes. *Genome Res* **19**, 1809–1816.

Park, D., Stanton, E., Ciezki, K., Parrell, D., Bozile, M., Pike, D., Forst, S. A., Jeong, K. C., Ivanek, R. & other authors (2013). Evolution of the Stx2-encoding prophage in persistent bovine *Escherichia coli* O157:H7 strains. *Appl Environ Microbiol* **79**, 1563–1572.

Perna, N., Plunkett Iii, G., Burland, V., Mau, B., Glasner, J., Rose, D., Mayhew, G., Evans, P., Gregor, J. & other authors (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409**, 529–533.

Rice, P., Longden, I. & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite (2000). *Trends Genet* **16**, 276–277.

Scheutz, F., Teel, L. D., Beutin, L., Piérard, D., Buvens, G., Karch, H., Mellmann, A., Caprioli, A., Tozzoli, R. & other authors (2012). Multicenter evaluation of a sequence-based protocol for subtyping Shiga toxins and standardizing Stx nomenclature. *J Clin Microbiol* **50**, 2951–2963.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069.

Shaikh, N. & Tarr, P. I. (2003). *Escherichia coli* O157:H7 Shiga toxin-encoding bacteriophages: integrations, excisions, truncations, and evolutionary implications. *J Bacteriol* **185**, 3596–3605.

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. (2006). ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* **34**, D32–D36.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313.

Stewart, A. C., Osborne, B. & Read, T. D. (2009). DIYA: a bacterial annotation pipeline for any genomics lab. *Bioinformatics* **25**, 962–963.

Sullivan, M. J., Petty, N. K. & Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics* **27**, 1009–1010.

Tarr, P., Gordon, C. & Chandler, W. (2005). Shiga-toxin-producing *Escherichia coli* and haemolytic uraemic syndrome. *Lancet* **365**, 1073–1086.

Tobe, T., Beatson, S. A., Taniguchi, H., Abe, H., Bailey, C. M., Fivian, A., Younis, R., Matthews, S., Marches, O. & other authors (2006). An extensive repertoire of type III secretion effectors in *Escherichia coli* O157 and the role of lambdoid phages in their dissemination. *Proc Natl Acad Sci U S A* **103**, 14941–14946.

Toro, M., Rump, L. V., Cao, G., Meng, J., Brown, E. W. & Gonzalez-Escalona, N. (2015). Simultaneous presence of insertion sequence excision enhancer and insertion sequence IS629 correlates with increased diversity and virulence in Shiga toxin-producing *Escherichia coli*. *J Clin Microbiol* **53**, 3466–3473.

Tree, J. J., Wolfson, E. B., Wang, D., Roe, A. J. & Gally, D. L. (2009). Controlling injection: regulation of type III secretion in enterohaemorrhagic *Escherichia coli*. *Trends Microbiol* **17**, 361–370.

Tyler, J. S., Beeri, K., Reynolds, J. L., Alteri, C. J., Skinner, K. G., Friedman, J. H., Eaton, K. A. & Friedman, D. I. (2013). Prophage induction is enhanced and required for renal disease and lethality in an EHEC mouse model. *PLoS Pathog* **9**, e1003236.

Wickham, H. (2009). *Ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag.

Yin, S., Rusconi, B., Sanjar, F., Goswami, K., Xiaoli, L., Eppinger, M. & Dudley, E. G. (2015). *Escherichia coli* O157:H7 strains harbor at least three distinct sequence types of Shiga toxin 2a-converting phages. *BMC Genomics* **16**, 733.

Zhang, Y., Laing, C., Steele, M., Ziebell, K., Johnson, R., Benson, A. K., Taboada, E. & Gannon, V. P. J. (2007). Genome evolution in major *Escherichia coli* O157:H7 lineages. *BMC Genomics* **8**, 121.

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. & Wishart, D. S. (2011). PHAST: a fast phage search tool. *Nucleic Acids Res* **39**, W347–W352.

Data Bibliography

Brittnacher, M., Jacobs, M., Zhou, Y., Chang, J., Fong, C., Gillett, W., Haugen, E., Hayden, H., Kulasekara, B., Larson Freeman, T., Radey, M., Rohmer, L., Sims, E., Wu, Z., Whittam, T., Kaul, R., Olson, M. V. & Miller, S. I. NCBI Reference Sequence NC_013008 (2016).

Cowley, C. A. GenBank CP015832 (2016).

Dallman, T. J., Byrne, L., Ahston, P. M., Cowley, L. A., Perry, N. T., Petrovska L., Ellis, R. J., Elson, R., Underwood, A., Green, J., Hanage, W. P., Jenkins, C., Grant, K. & Wain, J. BioProject PRJNA248042 (2015).

Eppinger, M., Sebastian, Y. & Ravel, J. NCBI Reference Sequence NC_011353 (2016).

Katani, R., Cote, R., Raygoza Garay, J. A., Li, L., Arthur, T. M., DebRoy, C., Mwangi, M. M. & Kapur, V. GenBank CP010304 (2014).

Latif, H., Aziz, R. K., Charusanti, P. & Palsson, B. O. GenBank CP008957 (2014).

Makino, K., Yokoyama, K., Kubota, Y., Yutsudo, C. H., Kimura, S., Kurokawa, K., Ishii, K., Hattori, M., Tatsuno, I., Abe, H., Iida, T., Yamamoto, K., Onishi, M., Hayashi, T., Yasunaga, T., Honda, T., Sasakawa, C. & Shinagawa, H. NCBI Reference Sequence NC_002695 (2016).

2.3 Concluding Remarks

While I stand by the main approaches and basic conclusions presented in the published manuscript, in retrospect it is evident that the work presented in this paper was somewhat naïve. This paper attempts to delve into the complexity of prophage evolution both in the context of interactions with the bacterial genetic content as well as with other prophages. As stated in the discussion of the paper, this type of works requires a much larger number of genomes to give greater confidence about primary conclusions and yield further insight into the prophage evolution and fixation within the bacterial genome. However, since this paper was published more long-read generated genomes have become available in public databases, as well as through our continuing collaborations, and these were then used to investigate further findings presented in this paper. This extra data allowed me to divide the data presented in this paper into two chapters: (1) individual prophage content and how this can supplement outbreak tracking (**Chapter 3**), and (2) whole strain prophage content, and how it may play a role in the evolution of isolates (**Chapter 4**). These two points were the cornerstones of the rest of my PhD project, and are presented in the next two results chapters rather than being provided as a very large extension to this chapter. The research also allowed for a better understanding of some of the observations made in published work, such as the different sub-populations of certain prophages, and what at the time we called IS phage “entrapment”. While, this might still be a plausible explanation, I did not pursue the IS “entrapment” theory as the research took a turn towards looking at much larger genome rearrangements, indicating that IS activity may only be a small component of recombinational events driving genome plasticity in individual isolates of *E. coli* O157.

3 Further Investigation of Shiga Toxin Encoding Prophage Identity and Similarity

3.1 Introduction

3.1.1 Rationale and Aim

This chapter holds a lot of similarities to **Chapter 2**, both in methodology, type of analysis, and target outputs. However, further sequencing was conducted to further investigate questions that were raised in Chapter 2, while also addressing limitation of the prior analysis.

As introduced in **Chapter 2**, nine isolates were first sequenced using PacBio. This was supplemented by 22 isolate sequences from Dr. Jim Bono at the USDA which he had previously sequenced. These 22 isolates were included in the analysis to provide a wider international genomic background, and provide a better frame of reference when comparing genomes of varying origins. Further sequencing of 33 UK isolates by PacBio was conducted by Dr. Jim Bono (USDA). These 33 isolates were selected for the following reasons:

1. There was need to have a large number of prophages from isolates of the same PT for comparison (as this was a limitation discussed in the previous chapter). PT 21/28 was selected as strain 9000 (**Table 3.1**) had an atypical Stx-encoding prophage, as discussed in **Chapter 2**, and the fact that PT 21/28 is one of the most prevalent PT causing human infection in the UK (**Chapter 3**).
2. Isolate 9000 was used by the Gally Lab for a couple of animal trials, therefore isolates from these trials were also sequenced to provide further background genomics of PT 21/28, but to also see the effect of passaging the strain through an animal (**Chapter 4**).
3. To investigate the differences in PFGE profiles (introduced in **Section 1.4**) amongst strains with a similar core SNP phylogeny (**Chapter 4**).

4. To investigate the differences between strains isolated at the same farm (**Chapter 4**).

Table 3.1 list all the sequences, their names, their source of origin, PT, and whether they were selected for one of the specific aims list above. **Table 3.1** also includes three sequences (one complete, and two partial), provided by PHE that were obtained using Oxford Nanopore sequencing.

Due to time constraints and new observations the prophage-centric analysis (looking at individual prophages) presented in **Chapter 2** had to be focused solely on Stx-encoding prophages for **Chapter 3**.

3.1.2 Long-Read Sequencing and Prophages

Long-read sequencing is introduced in the manuscript within **Chapters 1** and **2**, however, this section will review some of the points discussed in **Chapter 1** and expand on some statements made in **Chapter 2**. Two different types of long-read sequencing were utilised for this section: Pacific Bioscience (PacBio) sequencing (Pacific Biosciences 2019), and Oxford Nanopore MinIon sequencing (Oxford Nanopore 2019). The difference in chemistry between these two methods and short-read sequencing from Illumina has been discussed in **Chapter 1**. However, the output between the two long-read sequencing platforms is not identical. In **Chapter 2** it is mentioned that through the work of Dr. Jim Bono most of our sequences were assembled with the chromosomes spanning single contigs. This was partly due to the technology, but also Dr. Bono's experience with the platform, and manual curation to fully close certain genomes. The MinIon platform, on the other hand, is relatively new. As such, it is much less consistent, and while it shows great promise, due to its lower throughput, the initial strains sequenced by MinIon at PHE were unable to be closed as single contigs (**Table 3.1**). This is likely to be due to library preparation methods which are still evolving. While the outputs were multi-contig assemblies, Stx-encoding prophages tended to be fully assembled as part of larger contigs, and these could be used for further analysis. This offers an important opportunity, for as MinIon

sequencing improves, it can become more common for prophage sequences to be generated and analysed as part of PH responses to outbreaks. This can already be seen at PHE, where they have started to MinIon sequence any STEC of interest, in order to obtain further information after core phylogeny has been conducted using short-read sequencing. This novel advance will be further discussed in **Section 3.4.1**.

3.1.3 Sequence Alignments

Sequence alignments and pattern searching have been a staple of Bioinformatics for a long time. These are the main tools allowing for the comparison of sequences. Alignments are near ubiquitous in many tools and can be performed using a variety of algorithms. Read mapping mentioned in **Chapter 1** is the alignment of reads to a reference sequence, assemblies are the alignment of reads to one another, BLAST (Altschul et al. 1990; Camacho et al. 2009), a key Bioinformatics tool, stands for Basic Local Alignment Search Tools, and even k-mer identifying, simply streamlines the process by aligning short bits of sequence. As such, while alignments may seem mundane, they remain a very powerful tool which was heavily used in this analysis.

For this section we will focus on the mechanism behind BLAST, as most of the alignments from this point forward will either be standard BLAST results, or a derivation of it. The algorithm behind BLAST centers on a rapid variation of the Smith-Waterman Algorithm (SWA) (Smith and Waterman 1981) (which itself is a variation of the Needleman-Wunsch Algorithm (NWA) (Needleman and Wunsch 1970)). The NWA (Needleman and Wunsch 1970) uses dynamic programming, meaning that it breaks a large problem into smaller problems of the same type. The algorithm aims to obtain the optimal alignment between two sequences, in a way that is faster than trying all possible alignments, which would be extremely time and compute intensive. To do this a scoring matrix is generated and scores each base individually dependent on whether it matches, mismatches, or there is a gap between the

two sequences. Once the scoring matrix is generated, a traceback is initiated to determine the highest scoring path across the matrix, representing the best match. This method greatly increases the speed of alignments, while guaranteeing an optimal alignment is found. The SWA (Smith and Waterman 1981) is a variation of the NWA, where negative scores are not allowed in the matrix, with any negative cell being set to zero. While this may sound like a small difference, it allows for local alignments (alignments that do not involve the complete query sequences) by stopping the traceback when it encounters a cell with a score of zero. This can rapidly increase the speed of the algorithm, while also allowing for the search of smaller query sequences in a larger reference sequence.

While these two previous algorithms aim to determine the optimal alignment, BLAST (Altschul et al. 1990; Camacho et al. 2009) uses a heuristic approach to rapidly detect statistically significant alignments. This is done using seeding and extension. Seeding refers to the creation of seeds, or a small set of the query sequence, which are then searched for in the reference sequences. Each query sequence is divided into all the potential seeds within it, and once detected in the reference sequence, this becomes the search space that the algorithm will investigate. This allows rapid reduction of the space that the algorithm needs to parse. To further that reduction, the seeds are then extended left and right with an overall score being given to the match. Once the score drops below a certain threshold, the match is thought of as completed, and is then evaluated against all the other matches to determine whether its score is statistically significant or not. While the core local alignment algorithm resembles the SWA, it improves on it greatly due to the tremendous amount of data it can parse rapidly.

3.1.4 PHAST, PHASTER, and Prokka

The PHage Search Tool (PHAST) (Zhou et al. 2011) was the main online-server tool previously used for prophage identification and isolation. Other methods relying on BLAST or manual curation were used originally. While

these were more precise, they were also a lot slower and much less exhaustive but allowed us to validate and gauge the results obtained from PHAST (further discussed in **Section 3.2.2**). However, between the publication of the manuscript presented in **Chapter 2** and the arrival of all the PacBio sequences, PHage Search Tool – Enhanced Release (PHASTER) (Arndt et al. 2016) was released as a successor to PHAST. In this section, I will be describing how these tools differ, why a specific one was used over another, and the potential consequences these choices had.

The mechanism behind PHAST (Zhou et al. 2011) can be divided in two key sections: prophage gene detection, and gene density detection. Prophage gene detection was mostly done by combining three key methods. The first one was gene prediction and translation using the Open Reading Frame (ORF) predictor tool named GLIMMER (Delcher et al. 2007). These predicted genes and their translated proteins are then identified using BLAST to allow for a homology annotation. The final method consists of identifying phage sequence by use of BLAST to query a database of curated and annotated phage-specific sequences collected by the creators of PHAST. This allows for an exhaustive detection of phage associated genes across the query sequence. However, to detect prophages as a complete entity a cluster density analysis of these phage genes is conducted. This allows the user to determine whether there is a significantly higher concentration of phage genes within a specific region of the query sequence and call that region a prophage. Further analysis determines the likelihood the found match is a real prophage based on the length of the predicted prophage, and the presence of gene functional groups that are typical of a prophage. This method is effective at detecting prophages but has issues in clearly delineating the boundaries of the prophage regions, as shown by comparing the results to our other methods, and even to the newer version of the software named PHASTER.

PHASTER (Arndt et al. 2016) is the latest version of the PHAST tool from the same authors. PHASTER aims to improve on the process by having updated

its version of BLAST and having improved their hardware as well as optimized their code (other changes were made for metagenomic and multi-contig queries, but these did not apply to this body of work). A comparative run was conducted on PHASTER, and while a lower run time was achieved, boundaries varied slightly (with more overlapping prophage boundaries), and shorter prophages (which we considered prophage remnants in **Chapter 2**) showed the highest level of discrepancy in being called by the two versions (**Appendix II-A**). From this section onwards, PHASTER was used as its gene database is still being updated, unlike PHAST. To note, as described in **Chapter 2**, overlapping prophages were not merged as the mechanism behind PHAST and PHASTER would appear to indicate these to be discernible individual prophages.

Prokka: rapid prokaryotic genome annotation (Seemann 2014) was the software used for gene annotations within the prophages. It is a well-established annotator developed by T. Seemann and offers a great level of flexibility which was needed for this analysis. While the annotations obtained from Jim Bono using the DIY annotator (Stewart, Osborne, and Read 2009) (as mentioned previously) were of a high quality, prophage regions generally do not annotate well due to their versatility and general lack of reference genes within databases. Prokka (Seemann 2014) offers a function to create a gene database from genbank files. As such one would be able to make a reference gene database from EHEC O157 reference sequence annotations available on NCBI, the prophage gene database available on PHASTER, and any prior annotations from another pipeline (such as the DIY annotator pipeline). This not only allows for more robust naming of prophage associated genes but also more consistent naming of these genes. Prokka (Seemann 2014) also offers other features for a more specific annotation process, however, at its core uses similar tools and methods as other annotators with Prodigal (Hyatt et al. 2010) being the CDS caller, and BLAST the sequence match identifier.

3.2 Methods

All code written for the purpose of this chapter can be found at: https://github.com/SharifShaaban/PhD-Code (subfolders Appendix II-B and II-C). The following tool versions were used: EMBOSS 6.1.0, Perl 5.18.1 and 5.14.2, BLAST+ 2.4.0, Roary 3.6.0, R 3.2.2 and 3.3.1, Python 2.7.9, and Easyfig CL 2.1.

3.2.1 Genome Sequencing, Assembly and Annotation

Genome sequencing, assembly, and the original genome annotations were conducted by Dr. Jim Bono in the same way introduced in **Chapter 2**: using a combination of CANU (Koren et al. 2017), Geneious (Geneious 2019), Glimmer (Delcher et al. 2007), ORIFinder (Luo, Zhang, and Gao 2014) and DIY Annotator (Stewart, Osborne, and Read 2009). However, this time the number of sequenced strains was much larger, with a total of 69 EHEC O157:H7 isolates sequenced with the PacBio platform and having fully closed genomes, one isolate fully sequenced with the Minion platform and with its genome in multiple contigs (isolate PHE2), and two strains partially sequenced (only Stx2a-encoding prophages provided) by PHE. The 70 fully sequenced strains included 22 strains isolated from cattle obtained through Dr. Bono's affiliation with the USDA. Little metadata was made available from these strains, notably the lack of phage type as the US does not run that test. Therefore, these strains were viewed as a snapshot of the US cattle EHEC O157:H7 population. Five of the total strains were reference strains obtained from NCBI as presented in **Chapter 2** (Sakai, SS52, TW14359, EC4115, and EDL 933, respective accession numbers: BA000007, NZ_CP010304, NC_013008, NC_011353, and NC_002655). The remaining strains were all UK strains obtained through different projects. Nine of these were the original strains sequenced and discussed in **Chapter 2.** The remaining 34 sequences were a mixture of cattle and clinical UK strains, one of which was a MinIon sequence provided by PHE (sequence: PHE2). **Table 3.1** regroups all the

strains, their origin, phage type when known, and any other metadata made available.

3.2.2 Prophage and Shiga Toxin Calling

In **Section 3.1.4** PHAST and PHASTER were discussed. However, prior to using them for prophage calling other methods were used. Two main methods showed reasonable success. One consisted of searching for known and databased prophage sequences from reference EHEC O157:H7 strains (such as Sakai, TW14359, and EDL 933) using BLAST. This method was very accurate, determining prophage boundaries precisely as these regions had been heavily curated and studied previously. However, as one would expect the drawback was that this method would not detect novel prophage sequences or only partially determine them. To resolve this for Stx-encoding prophages, manual curation was used. This method involved using BLAST to determine the location of Stx genes (using the (Scheutz et al. 2012) reference sequences), and then manually determine where the prophage associated gene annotations ended. This allowed for novel Stx prophages to be detected, and relatively accurate boundaries to be determined. However, this method was extremely time consuming (30-60 minutes per prophage), heavily relied on correct and complete gene annotation (which was not always available for prophage associated genes), and required an inordinate amount of manual curation. PHAST and PHASTER provided a balance between run-time (5-30 minutes to detect all the prophages within an EHEC strain), sensitivity, and specificity (further discussed in **Section 5.2**). There is one main differences in the prophage calling methodology between the results presented in **Chapter 2** and those in **Chapters 3** and **4**: PHASTER was used instead of PHAST for **Chapters 3** and **4**. Sequences were submitted using the PHASTER URLAPI, and prophages were extracted from the summary table using similar scripts as presented in **Chapter 2**; these can be found in **Appendix II-B**.

Stx-encoding prophages were determined using BLAST across the extracted prophage sequences. They were then clustered based on their Stx subtype, using reference sequences (Scheutz et al. 2012). In **Chapter 1** I discussed the difficulty of detecting Stx genes within EHEC O157:H7, however, long-read sequencing mostly negates this by allowing for full genomes to be assembled within single contigs, and multiple occurring Stx genes within single strains to be differentiated and assembled correctly. The script can be found in **Appendix II-C**.

Name	Country	Source	Provider	Phage Type	Notes	1a	2a	2c	2d
16438	U.K	Cattle	IPRAVE	32	None				
Z1486	U.K	Cattle	IPRAVE	21/28	None				
Z1504	U.K	Cattle	IPRAVE	21/28	None				
Z1811	U.K	Cattle	IPRAVE	21/28	Profile C (11b)				
Z563	U.K	Cattle	IPRAVE	21/28	None				
Z570	U.K	Cattle	IPRAVE	21/28	None				
Z852	U.K	Cattle	IPRAVE	21/28	None				
Z866	U.K	Cattle	IPRAVE	21/28	None		2		
Z887	U.K	Cattle	IPRAVE	21/28	None				
Z892	U.K	Cattle	IPRAVE	21/28	None				
Z903	U.K	Cattle	IPRAVE	21/28	None				
Z910	U.K	Cattle	IPRAVE	21/28	None				
7784	U.K	Cattle	IPRAVE	32	None				
Z1814	U.K	Cattle	IPRAVE	21/28	None				
EC4115	U.S.A	Cattle	NCBI	N/A	Accession: NC_011353				
EDL933	U.S.A	Cattle	NCBI	N/A	Accession: NC_002655				
Sakai	Japan	Human	NCBI	N/A	Accession: BA000007				
SS52	U.S.A	Cattle	NCBI	N/A	Supershedder, Accession: NZ_CP010304				
TW14359	U.S.A	Cattle	NCBI	N/A	Accession: NC_013008				
PHE2	U.K	Human	PHE	N/A	MinIon Sequenced, Multi-contig				
PHEO26	U.K	Human	PHE	N/A	O26, Stx2a, Multi-contig				
155	U.K	Human	PHE	32	Linked with Ireland				
180	U.K	Human	PHE	54	None	2			
272	U.K	Human	PHE	2	None				
319	U.K	Human	PHE	UT	None				
350	U.K	Human	PHE	8	None				
472	U.K	Human	PHE	14	None				
644	U.K	Human	PHE	8	None	2			
122262	U.K	Human	PHE	N/A	O55, Stx2a, Multi-contig				
Z1626	U.K	Human	SERL	21/28	Profile C (11b)				
Z1812	U.K	Human	SERL	21/28	Profile C (11b)				
Z1815	U.K	Human	SERL	21/28	None				
Z1816	U.K	Human	SERL	21/28	None				
Z1825	U.K	Human	SERL	21/28	None				
Z1826	U.K	Human	SERL	21/28	None				
Z1830	U.K	Human	SERL	21/28	Single Farm S101				
Z1831	U.K	Human	SERL	21/28	Single Farm S101				
Z1832	U.K	Human	SERL	21/28	Single Farm S101				
Z1833	U.K	Human	SERL	21/28	Single Farm S101				
Z1834	U.K	Human	SERL	8	Single Farm E018				
Z1835	U.K	Human	SERL	54	Single Farm E018				

Z1836	U.K	Human	SERL	54	Single Farm E018				
Z1813	U.K	Human	SERL	21/28	None				
U17B6	U.S.A	Cattle	USDA	N/A	None				
UBB24	U.S.A	Cattle	USDA	N/A	None				
UF6294	U.S.A	Cattle	USDA	N/A	None				
UF6667	U.S.A	Cattle	USDA	N/A	None				
UF7386	U.S.A	Cattle	USDA	N/A	None				
UF7508	U.S.A	Cattle	USDA	N/A	None				
UF8797	U.S.A	Cattle	USDA	N/A	None				
UF8952	U.S.A	Cattle	USDA	N/A	None				
UGI11	U.S.A	Cattle	USDA	N/A	None				
UGI351	U.S.A	Cattle	USDA	N/A	None				
UH2495	U.S.A	Cattle	USDA	N/A	None				
UKS470	U.S.A	Cattle	USDA	N/A	None				
UMB41	U.S.A	Cattle	USDA	N/A	None				
UN8B7	U.S.A	Cattle	USDA	N/A	None				
UTB21	U.S.A	Cattle	USDA	N/A	None				
UTX265	U.S.A	Cattle	USDA	N/A	None				
UTX313	U.S.A	Cattle	USDA	N/A	None				
UTX754	U.S.A	Cattle	USDA	N/A	None				
UU44	U.S.A	Cattle	USDA	N/A	None				
UU78	U.S.A	Cattle	USDA	N/A	None				
UU87	U.S.A	Cattle	USDA	N/A	None				
UYB14	U.S.A	Cattle	USDA	N/A	None				
Z1615	U.K*	Cattle*	ZAP	21/28*	Trial -High Shedder #1				
Z1766	U.K*	Cattle*	ZAP	21/28*	Trial -isolate variation				
Z1767	U.K*	Cattle*	ZAP	21/28*	Trial -isolate variation				
Z1768	U.K*	Cattle*	ZAP	21/28*	Trial -isolate variation				
Z1769	U.K*	Cattle*	ZAP	21/28*	Trial -isolate variation				
9000	U.K	Cattle	IPRAVE	21/28	Stx2a gene disrupted by IS				
10671	U.K	Cattle	IPRAVE	32	None				

Table 3.1 List of all the strains used in **Chapters 3 and 4**, their country of origin, the source they were isolated from, the lab that provided the isolates (ZAP being the designation for strains stored in David Gally's laboratory), their phage type, and relevant comments about the strains. Strains Z1615, Z1766-1769 contain asterisks in their country, source and phage type, as while these are correct, these strains were generated through a project inoculating cattle with strain 9000 (indicated in Notes as Trial). The Note titled Single Farm indicates that these samples were isolated within individual farm in order to sample the diversity. Notes indicating Profile C (11b) point to the PFGE profile of these strains. Profile C (11b) was a PFGE profile detected by SERL which remained consistently present across time within clinical isolates. Isolates PHEO26 and 122262 only had their Stx2a prophage sequences provided by PHE. Stx presence and absence within these genomes is also displayed with green indicating the presence of this Stx subtype (number within cell represents multiple copies), and red indicates absence of the subtype (determined using BLAST).

3.2.3 Prophage Clustering

For this chapter not all prophages were clustered as was previously carried out in **Chapter 2**. Only Stx prophages were clustered based on their predicted subtypes which was determined by their highest BLAST match score to the reference *stx* sequences. Similarly to **Chapter 2**, prophage genes were re-annotated using Prokka and a prophage gene database generated as described in **Section 3.1.4**; this was coloured by functional groups to simplify visualisation of the comparison. Time constraints meant foregoing the clustering of all prophages and focusing on Stx-encoding prophages (**Chapter 3**) and the data presented in **Chapter 4**, which offers a different hypothesis to the observations made in **Chapter 2**. An opportunity for further work exists as all the scripts were already written to automatically run and results still generated (**Appendix II-B)** therefore analysis should be possible.

Stx-encoding prophages of different subtypes were separated and sub-clustered as well. It follows a similar method as defined in **Chapter 2**, however, this sub-clustering was performed manually based on EasyFig alignments (Sullivan, Petty, and Beatson 2011). Prophages exhibiting no changes in gene content were clustered together. IS element movement and duplication were allowed and did not result in samples being clustered separately.

3.2.4 Prophage Alignments and Phylogenic Investigation

The protocol here was identical to that applied in **Chapter 2**. Prophage alignment was conducted using EasyFig (Sullivan, Petty, and Beatson 2011), a BLAST result visualizer. It was run only looking for matches larger than 1200 bp to avoid smaller areas of homology within genes being displayed, this make the figures easier to interpret and IS movement is generally not depicted. Genes were displayed as arrows and coloured based on their functional groups, with whole IS elements delineated as smaller yellow frames. The order of prophages in the figure was determined through a

manual iterative process where the alignment was run, similar prophages were ordered adjacent to one another, and the alignment run once more, until prophage were sorted alongside similar prophages.

Based on these alignment (**Figures 3.3**, **3.4**, and **3.5**) prophage groupings were determined based on genetic content as follows:

- Prophages exhibiting the same genetic content except for IS changes were classified in the same group and subgroup
- Prophages exhibiting only small differences within genes were classified in the same group
- Groups are represented using a Greek alphabet letter
- Subgroups are represented by a number superscripted over the group
- If all members of a group belong to the same subgroup, no superscript subgroup is represented
- An asterisk next to the group name indicates that prophages belonging to that group or subgroup showed 100% identity and coverage with no differences
- Group θ was created for prophages that could not be grouped with the rest.
- Prophage groupings are not comparable between the different Stx-encoding prophage types (e.g. a Stx1a-encoding prophage of group α is not identical to a Stx2a-encoding prophage of group α)

The phylogenetic investigation was conducted on Stx-encoding prophages based on the hypothesis that strains with near identical prophages might have a similar origin. Prophage similarity and identity was tested by observing the EasyFig alignments as well as further delving within the BLAST temporary files generated by EasyFig. These files are a typical BLAST output file, and therefore indicate the exact percentage coverage and identity that

two prophages share. Furthermore, as previously stated an aim of this chapter was to investigate the usefulness of such prophage / long-read sequencing data for PH action. As such PHE provided us with Stx-encoding prophage sequences that they generated through MinIon sequencing. These sequences came with location metadata which allowed for certain observations to be made, and hypothesis to be generated towards the origin of these strains, and how the prophage population might help supplement core phylogeny (**Section 3.4.1**).

3.3 Results

3.3.1 Prophage Calling

The following results until **Section 3.3.3** only discuss prophages obtained from the 70 isolates with complete genome sequences (69 isolates in single-contig assemblies and sequence PHE2) and omits the separate Stx-encoding prophage sequences supplied by PHE (as these were not full genome assemblies). A total of 1163 prophages were detected in the 70 isolates ranging in size from 5580bp to 160293bp, these include prophages that were merged as described in **Chapter 2** and **Section 3.1.4**. **Figure 3.1** regroups the frequency of prophage lengths observed. On average 17 prophages (rounded to the nearest integer) were detected per isolate (minimum: 14, maximum: 21 prophages). **Table 3.2** contains the number of prophages detected within each isolate.

3.3.2 Prophage Similarity

While in **Chapter 3** the relevance of Sakai prophages as a reference within a more varied population of strains is questioned, a comparison analysis was still conducted using a similar method. **Table 3.2** contains a heat map of the results found. As in **Chapter 2**, only a small subset of the Sakai prophages was present at a high level of similarity within a majority of the isolates (mainly SP3, SP13, and SP14), and therefore their use as “reference” prophages may not be that valuable as they are not representative of the worldwide EHEC O157 prophage population.

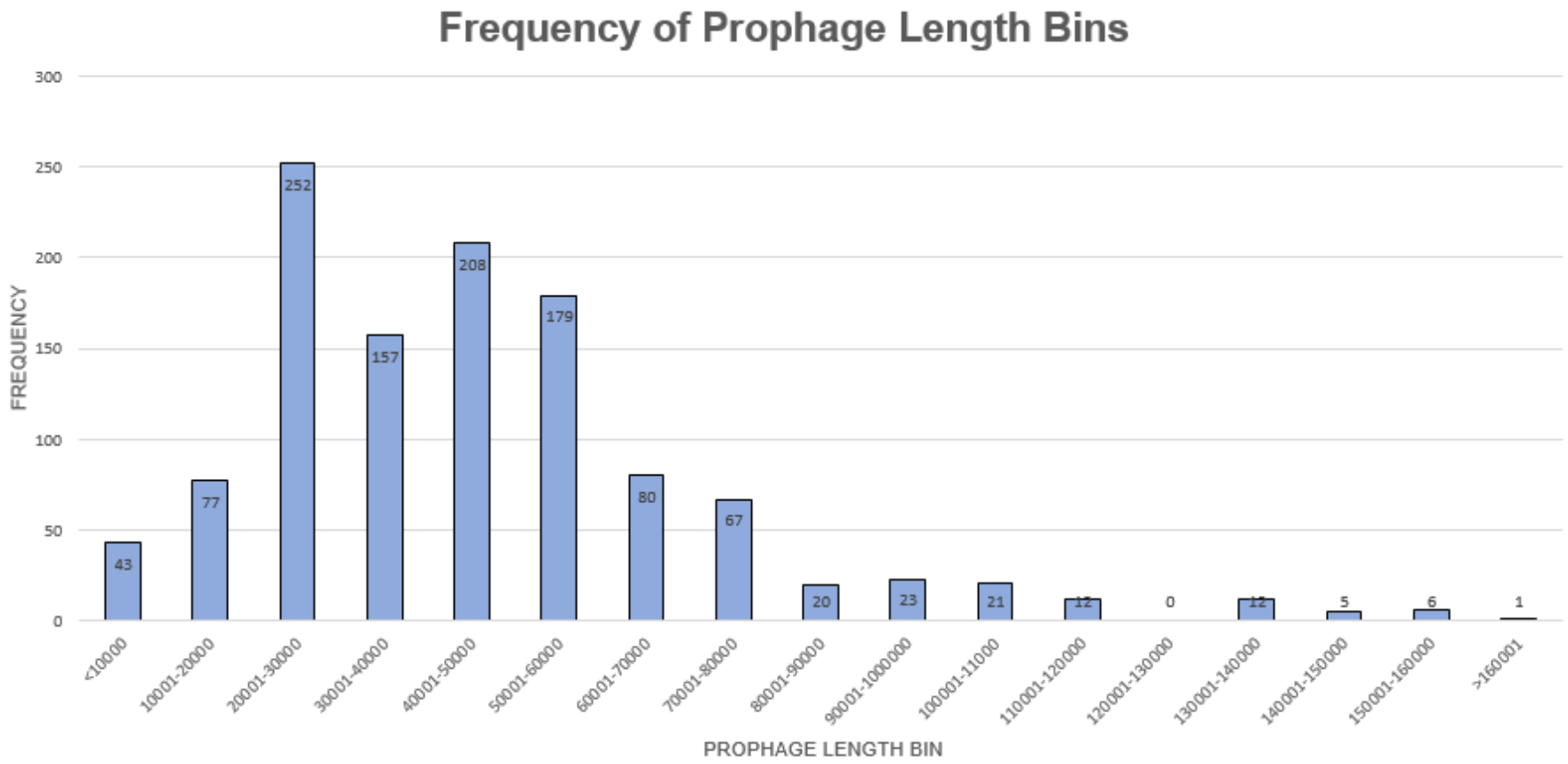


Figure 3.1 Diagram reprenting the frequencies of different prophage lengths detected within the 70 whole genome sequences of the isolates subdivided in length ranges of 10000 bp.

Isolate	Number of Prophages	SP 1,2	SP 3	SP 4	SP 5	SP 6	SP 7	SP 8	SP 9	SP 10	SP 11,12	SP 13	SP 14	SP 15	SP 16*	SP 17	SP 18*
16438	17																
Z1486	16																
Z1504	16																
Z1811	17																
Z563	16																
Z570	17																
Z852	16																
Z866	17																
Z887	16																
Z892	16																
Z903	16																
Z910	15																
7784	15																
Z1814	17																
EC4115	17																
EDL933	14																
Sakai	17																
SS52	16																
TW14359	17																
PHE2	17																
155	18																
180	15																
272	16																
319	16																
350	16																
472	17																
644	18																
Z1626	16																
Z1812	16																
Z1815	15																
Z1816	15																
Z1825	17																
Z1826	16																
Z1830	16																
Z1831	16																
Z1832	16																
Z1833	16																
Z1834	17																
Z1835	17																
Z1836	17																
Z1813	16																
U17B6	17																
UBB24	20																
UF6294	16																
UF6667	17																
UF7386	15																
UF7508	16																
UF8797	16																
UF8952	14																
UGI11	14																
UGI351	16																
UH2495	16																
UKS470	18																
UMB41	18																
UN8B7	17																
UTB21	18																
UTX265	16																
UTX313	21																
UTX754	19																
UU44	16																
UU78	20																
UU87	19																
UYB14	17																
Z1615	16																
Z1766	17																
Z1767	18																
Z1768	17																
Z1769	17																
9000	17																
10671	18																

Table 3.2 List of all the 70 isolates which had a whole genome sequence, the number of prophages detected within them, and a colour code indicated whether a SP was detected within these sequences with near perfect identity or high similarity.
Legend: **Similar prophage at a distance of t0 detected (indicating identical gene content but potentially different gene order or number of gene duplicates. See Chapter 3).**
Similar prophage at a distance of t4.5 detected (approximating 80% coverage and identity across the whole prophage. See Chapter 3).
No prophage at a distance below t4.5 detected.
* Multiple prophages within single strains exhibited levels of similarities to these SPs.

3.3.3 Shiga Toxin Encoding Prophages

In total 125 Stx-encoding prophages were detected, including the two Stx2a prophages provided by PHE. However, only 125 of these were used for further analysis as there was a sole Stx2d-encoding prophage (in isolate UF6294) and this could not be clustered. Stx2a, Stx2c, and Stx1a prophages were aligned with 52, 57, and 16 prophages belonging to each subtype respectively. It should be noted that one prophage had to be included in both the Stx2a and Stx2c clusters as it appeared to be two merged prophages with both Stx subtypes (from strain UU87). However, as we have limited metadata regarding that isolate and that it couldn't be confirmed whether this was an artifact of assembly or an actual prophage merge, this isolate while included in the complete alignments in **Appendix II-D**, was dismissed from further discussions.

Of the 16 Stx1a-encoding prophages, four sub-clusters (with one sequence being alone) were identified (**Figures 3.2** and **3.3**), the largest one (β) was found only in UK isolates and had completely identical sequences.

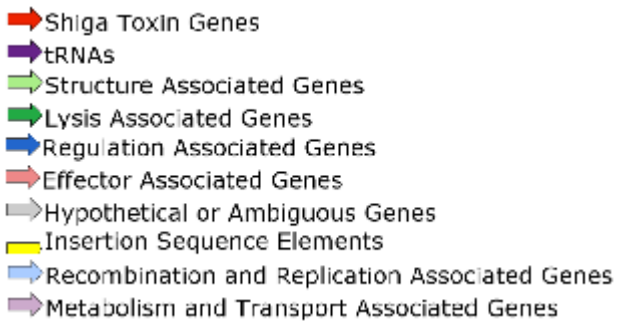


Figure 3.2 Gene colour legend for **Figures 3.3**, **3.4** and **3.5**.

Label	Name	Source	Phage Type	Coordinates	Group
1	**472**	Human	14	3922497-4000398	**θ**
2	**180**	Human	54	3881990-3952213	α*
3	**644**	Human	8	4199527-427581	α*
4	**UBB24**	Cattle	N/A	3762079-3825049	δ
5	**UTX265**	Cattle	N/A	3757957-3816461	δ
6	**180**	Human	54	3592886-3648312	β*
7	**319**	Human	UT	3614264-3668374	β*
8	**350**	Human	8	3582492-3637091	β*
9	**644**	Human	8	3915380-3969656	β*
10	**Z1834**	Human	8	3654221-3708496	β*
11	**Sakai**	Human	N/A	3668037-3725542	γ
12	**EDL933**	Cattle	N/A	3717605-3775110	γ
13	**UF6294**	Cattle	N/A	3767688-3820355	γ
14	**UF7508**	Cattle	N/A	3677321-3734090	γ
15	**UH2495**	Cattle	N/A	3667149-3723340	γ
16	**UU87**	Cattle	N/A	3737248-3793858	γ

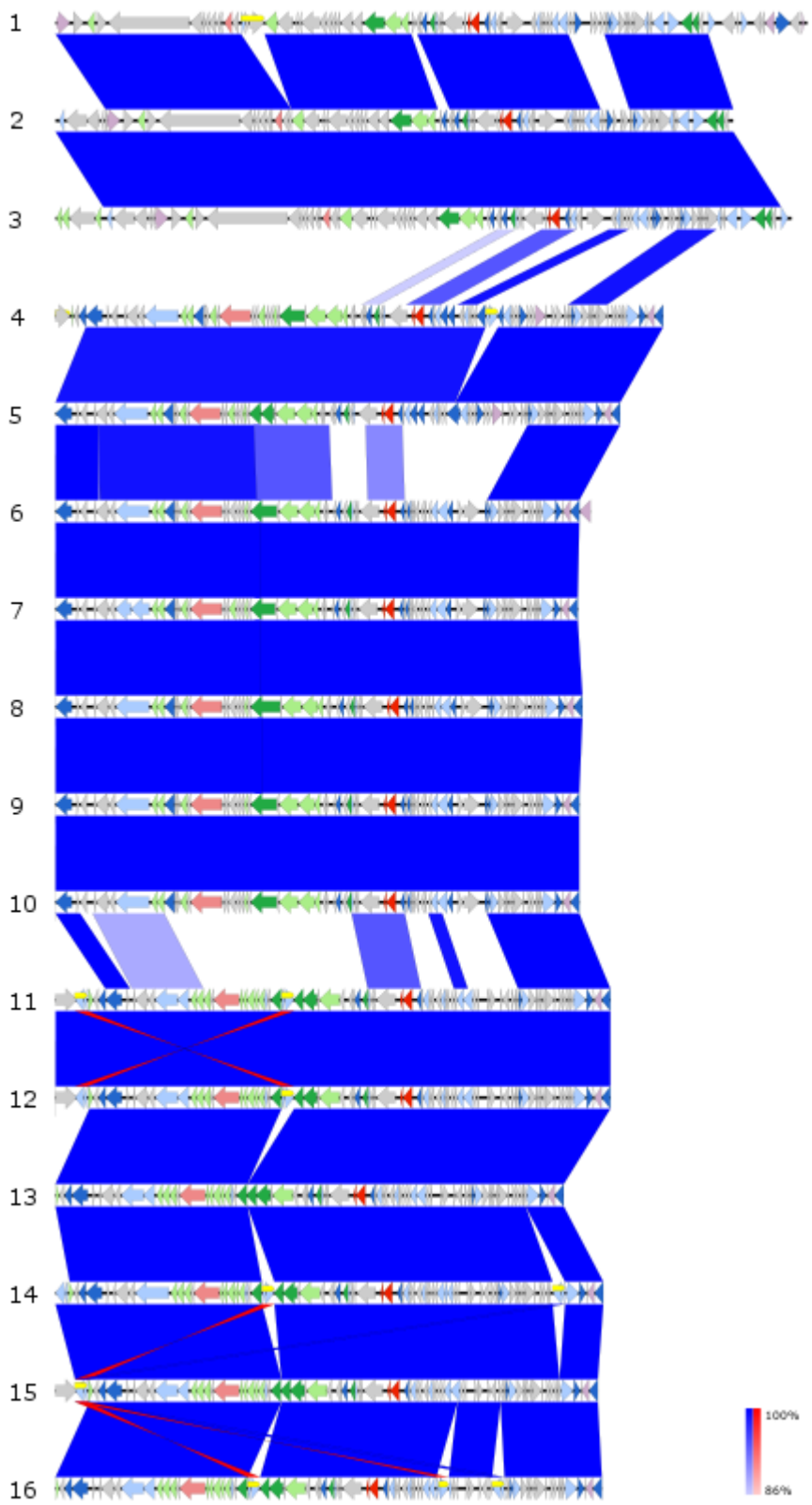


Figure 3.3 EasyFig alignment of Stx1a-encoding prophages. The name, source, and phage type of the strains is included in the table. The full alignment can be found in **Appendix II-D**. It would appear that most UK strains are within sub-cluster β which is highly conserved. Other groupings appear to be more diverse in their source and phage types, with no clear differentiation between strains from the U.S.A and the UK. However, it should be noted that the variation across all these groups and subgroups is relatively small with a minimum BLAST identity of 86% when a match is present, and mostly preserved regulatory regions. Gene colour legend in **Figure 3.2**.

Label	Name	Source	Phage Type	Coordinates	Group
1	**UU87***	Cattle	N/A	3485516-3604416	θ
2	**Z563**	**Cattle**	**21/28**	**3408094-3470882**	**α***
3	**Z1767**	Cattle*	21/28*	3422303-3485091	α*
4	**472**	**Human**	**14**	**3450080-3511541**	**β^1**
5	**7784**	Cattle	32	3348763-3414075	β^1
6	**EC4115**	Cattle	N/A	3407186-3469712	β^1
7	**SS52**	Cattle	N/A	3368403-3430928	β^1
8	**TW14359**	Cattle	N/A	3406185-3468711	β^1
9	**UF8797**	Cattle	N/A	3359432-3421957	β^1
10	**UYB14**	Cattle	N/A	3521081-3584981	β^1
11	**10671**	**Cattle**	**32**	**3357907-3421580**	**β^2**
12	**UMB41**	Cattle	N/A	3437557-3501419	β^2
13	**UGI11**	**Cattle**	**N/A**	**3405365-3469588**	**β^3**
14	**UTB21**	Cattle	N/A	3478728-3541453	β^3
15	**U17B6**	**Cattle**	**N/A**	**3509896-3580331**	**β^4**
16	**UU44**	Cattle	N/A	3453407-3525076	β^4
17	**9000**	**Cattle**	**21/28**	**3416723-3485094**	**γ***
18	**16438**	Cattle	32	3417865-3477412	γ*
19	**Z892**	Cattle	21/28	3419198-3484404	γ*
20	**Z1486**	Cattle	21/28	3419200-3487562	γ*
21	**Z1504**	Cattle	21/28	3407449-3475819	γ*
22	**Z1811**	Cattle	21/28	3415663-3484033	γ*
23	**Z570**	Cattle	21/28	3457346-3525717	γ*
24	**Z852**	Cattle	21/28	3360501-3428872	γ*
25	**Z887**	Cattle	21/28	3435733-3504103	γ*
26	**Z910**	Cattle	21/28	3404876-3473247	γ*
27	**Z1626**	Human	21/28	3417930-3486300	γ*
28	**Z1812**	Human	21/28	3417703-3486079	γ*
29	**Z1815**	Human	21/28	3461144-3529514	γ*
30	**Z1816**	Human	21/28	3459773-3528142	γ*
31	**Z1825**	Human	21/28	3449084-3517456	γ*
32	**Z1826**	Human	21/28	3426654-3495026	γ*
33	**Z1830**	Human	21/28	3421530-3489892	γ*
34	**Z1831**	Human	21/28	3421811-3490174	γ*
35	**Z1832**	Human	21/28	3421822-3490184	γ*
36	**Z1833**	Human	21/28	3421790-3490152	γ*
37	**Z1615**	Cattle*	21/28*	3416709-3485079	γ*
38	**Z1766**	Cattle*	21/28*	3416722-3485092	γ*
39	**Z1768**	Cattle*	21/28*	3416719-3485089	γ*
40	**Z1769**	Cattle*	21/28*	3416714-3485084	γ*
41	**Z1813**	Human	21/28	3491845-3560215	γ*
42	**Z1814**	Cattle	21/28	3420120-3488490	γ*
43	**UKS470**	**Cattle**	**N/A**	**3509277-3579848**	**δ^{1*}**
44	**UN8B7**	Cattle	N/A	3429974-3500543	δ^{1*}
45	**UTX313**	Cattle	N/A	3609168-3679696	δ^{1*}
46	**UTX754**	Cattle	N/A	3518721-3589290	δ^{1*}
47	**UU78**	Cattle	N/A	3475711-3546281	δ^{1*}
48	**Z1835**	**Human**	**54**	**3427500-3499506**	**δ^2**
49	**Z1836**	Human	54	3475052-3547058	δ^2
50	**UBB24**	**Cattle**	**N/A**	**3519723-3594050**	**δ^3**
51	**UTX265**	Cattle	N/A	3516216-3589249	δ^3
52	**Z1834**	**Human**	**8**	**3414522-3484703**	**δ^4**
53	**350**	**Human**	**8**	**3345346-3412985**	**δ^5**
54	**644**	**Human**	**8**	**3679887-3745861**	**δ^6**
55	**180**	Human	54	3359761-3423367	δ^6
56	**Z903**	**Cattle**	**21/28**	**3421641-3496120**	**θ**
57	**UF6667**	**Cattle**	**N/A**	**3375219-3455307**	**θ**

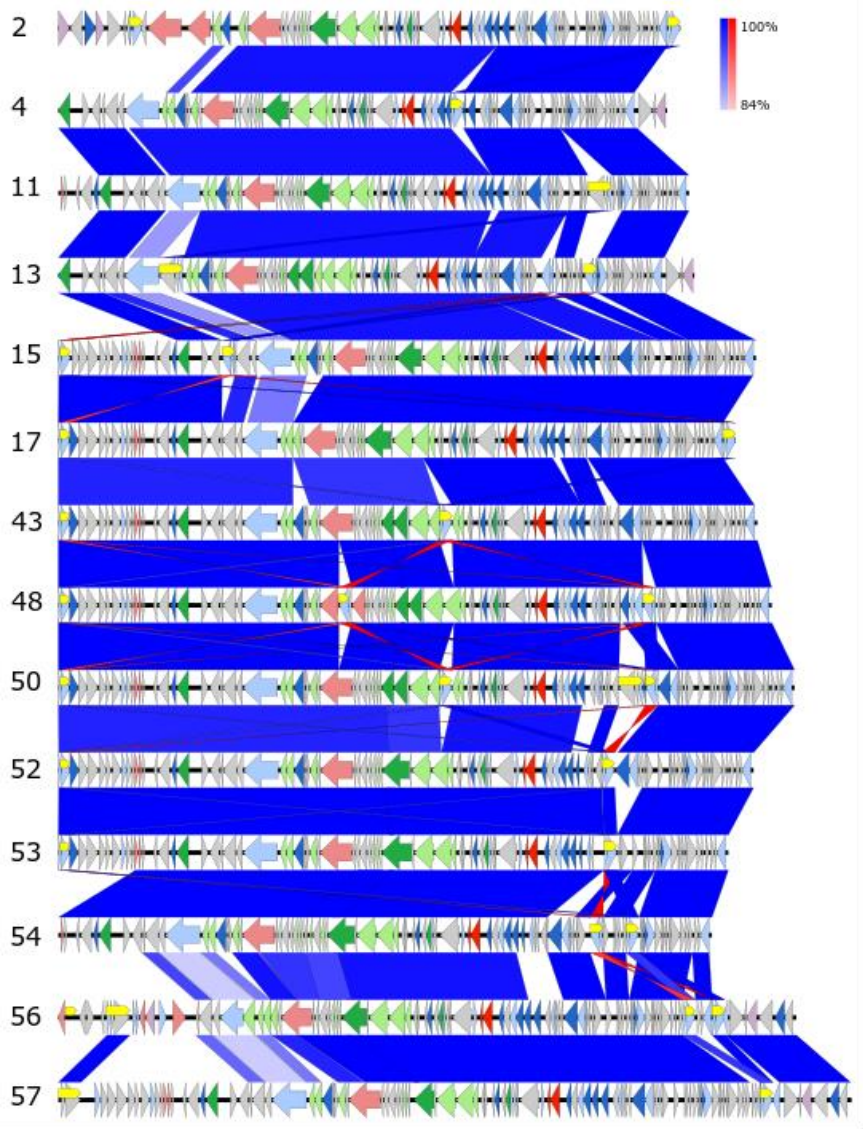


Figure 3.4 EasyFig alignment of Stx2c-encoding prophages. Due to their large numbers, each prophage group only has one representative prophage aligned. An asterisk next to the isolate name means that this prophage was dropped from further analysis. An asterisk next to either the source or PT in indicates that this strain is a derivative generated during a trial and, therefore, the PT is the suspected one. The name, source, and phage type of the strains is included in the table along with a label indicating the represented strains (in bold). The full alignment can be found in **Appendix II-D**. While Stx2c-encoding prophages were numerous, it is apparent that most strains with the phage type 21/28 contain an identical Stx2c-encoding prophage (group γ). Other groupings appear to be more diverse in their source and phage types, with no clear differentiation between strains from the U.S.A and the U.K except for phage type 21/28. However, it should be noted that the variation across all these groups and subgroups is relatively small with a minimum BLAST identity of 84% when a match is present, and mostly preserved regulatory regions. Gene colour legend in **Figure 3.2**.

Within the 57 Stx-2c encoding prophages (**Figures 3.2** and **3.4**) four sub-clusters were identified with the largest one (γ) exhibiting a 100% prophage identity across 26 sequences primarily of the PT 21/28 or expected to be. Including UU87, three Stx2c-encoding prophage sequences were not given clusters.

Six sub-clusters were identified within the 52 Stx2a-encoding prophage sequences (**Figures 3.2** and **3.5**). The largest cluster (γ) was once more representative of PT21/28 but exhibited more variation within it than its Stx2c counterpart. However, BLAST identity across all three comparisons ranged between 84% and 92% minimums when matches were found, indicating that variation was rarely due to mutation within genes or SNPs, but mainly to presence or absence of genes.

Of note is that overrepresentation of PT21/28 within the dataset rather may be the reason behind the fact that the γ clusters are the largest (**Figure 3.3** and **3.4**). Nonetheless the high consistency and minimal variation of these prophages within these clusters is worth noting and will be discussed further. Including UU87, four prophages were not assigned sub-cluster. Another prophage did not sub-cluster (Z866, 3463544-3530554) because while it was called a Stx2a-encoding prophage, it has the prophage content consistent with the Stx2c β sub-cluster. As this once more could not be verified whether real or a mis-assembly issue, this prophage will be dropped from further discussion. This can be seen in the complete alignments included in **Appendix II-D**, as **Figures 3.4** and **3.5** only show sub-cluster representatives as part of the alignments. The investigation of prophage Z866, 3463544-3530554 did lead to another observation: that Stx2a sub-cluster β appears to share a closer prophage background to Stx2c-encoding prophage than the rest of Stx2a-encoding prophages.

Label	Name	Source	Phage Type	Coordinates	Group
1	**Z866***	Cattle	21/28	3463544-3530554	stx2c
2	**UU87***	Cattle	N/A	3485516-3604416	Θ
3	**EC4115**	**Cattle**	**N/A**	**3974001-4051195**	**α***
4	**SS52**	Cattle	N/A	3889352-3966546	α*
5	**TW14359**	Cattle	N/A	3928890-4006123	α*
6	**UF7386**	Cattle	N/A	3805183-3882377	α*
7	**UF8797**	Cattle	N/A	3880825-3958019	α*
8	**Z1836**	**Human**	**54**	**3242249-3289469**	**β^1**
9	**PHEO26**	Human	N/A	N/A	β^1
10	**155**	**Human**	**32**	**3280940-3330611**	**β^2**
11	**PHE2**	Human	N/A	N/A	β^2
12	**122262**	**Human**	**N/A**	**N/A**	**β^3**
13	**9000**	**Cattle**	**21/28**	**3908289-3987569**	**γ**
14	**Z1486**	Cattle	21/28	3910757-3987589	γ
15	**Z1504**	Cattle	21/28	3899014-3979937	γ
16	**Z1811**	Cattle	21/28	3907228-3982741	γ
17	**Z563**	Cattle	21/28	3903570-3988737	γ
18	**Z570**	Cattle	21/28	3948912-4024469	γ
19	**Z852**	Cattle	21/28	3853378-3930210	γ
20	**Z866**	Cattle	21/28	3953797-4029355	γ
21	**Z887**	Cattle	21/28	3927303-4004174	γ
22	**Z892**	Cattle	21/28	3907580-3984412	γ
23	**Z903**	Cattle	21/28	3912097-3988929	γ
24	**Z910**	Cattle	21/28	3896442-3973274	γ
25	**Z1626**	Human	21/28	3909495-3986327	γ
26	**Z1812**	Human	21/28	3909276-3984794	γ
27	**Z1815**	Human	21/28	3952704-4030891	γ
28	**Z1816**	Human	21/28	3951336-4029522	γ
29	**Z1825**	Human	21/28	3940534-4017403	γ
30	**Z1826**	Human	21/28	3918221-3995050	γ
31	**Z1830**	Human	21/28	3913084-3989916	γ
32	**Z1831**	Human	21/28	3913366-3991511	γ
33	**Z1832**	Human	21/28	3913375-3991520	γ
34	**Z1833**	Human	21/28	3913342-3991486	γ
35	**Z1615**	Cattle*	21/28*	3917767-4002933	γ
36	**Z1766**	Cattle*	21/28*	3908287-3985117	γ
37	**Z1767**	Cattle*	21/28*	3908286-3985155	γ
38	**Z1768**	Cattle*	21/28*	3908283-3985113	γ
39	**Z1769**	Cattle*	21/28*	3908279-3985109	γ
40	**Z1813**	Human	21/28	3983414-4060285	γ
41	**Z1814**	Cattle	21/28	3911677-3988506	γ
42	**272**	**Human**	**2**	**3856341-3936140**	**δ**
43	**UGI351**	Cattle	N/A	3860112-3937285	δ
44	**Sakai**	**Human**	**N/A**	**2025166-2095660**	**ε**
45	**UF6294**	Cattle	N/A	1987550-2058044	ε
46	**UF8952**	Cattle	N/A	1986015-2057822	ε
47	**UH2495**	Cattle	N/A	2024959-2095453	ε
48	**EDL933**	**Cattle**	**N/A**	**2073887-2143336**	**ζ^1**
49	**UF7508**	**Cattle**	**N/A**	**2043406-2113125**	**ζ^2**
50	**16438**	**?**	**32**	**3911835-3989726**	**θ**
51	**UF6667**	**Cattle**	**N/A**	**3902119-3976877**	**θ**
52	**UU44**	**Cattle**	**N/A**	**3954327-4029564**	**θ**

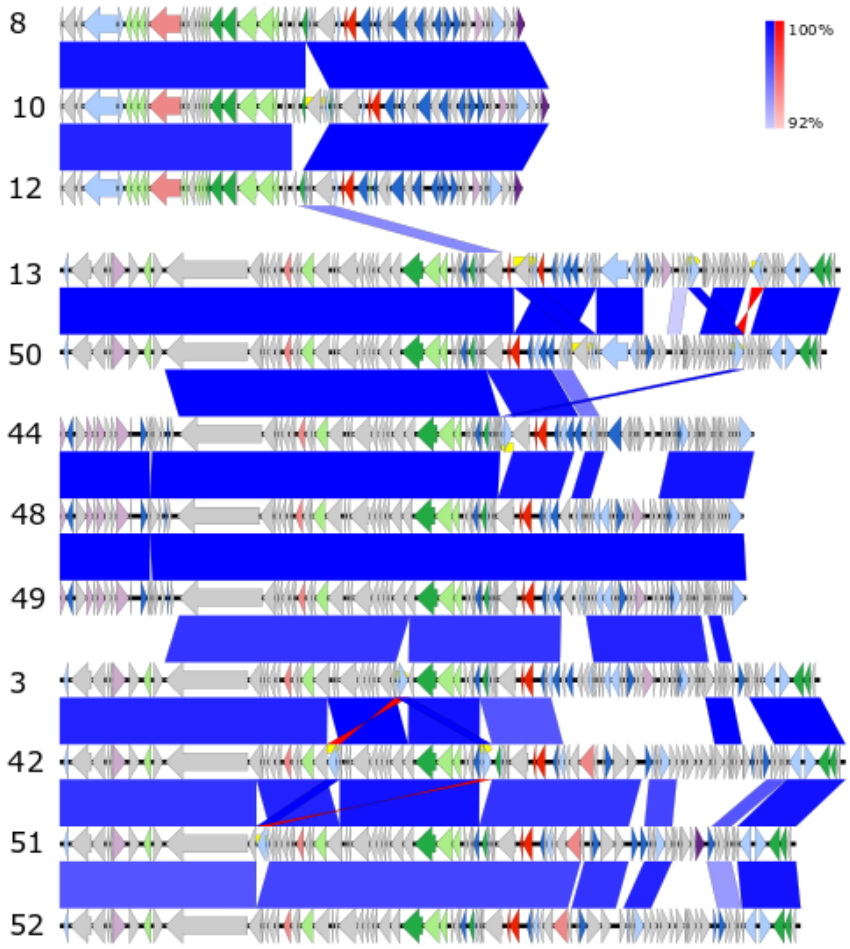


Figure 3.5 EasyFig alignment of Stx2a-encoding prophages. Due to their large numbers, each prophage group only has one representative prophage aligned. An asterisk next to the isolate name means this prophage was dropped from further analysis. An asterisk next to either the source or PT in indicates that this strain is a derivative generated during a trial and, therefore, the PT is the suspected one. The name, source, and phage type of the strains is included in the table along with a label indicating the represented strains (in bold). The full alignment can be found in **Appendix II-D**. While Stx2a-encoding prophages were numerous, it is apparent that most PT21/28 isolates contain a near identical Stx2a-encoding prophage (group γ). Only five U.K isolates were not within this group. All the prophages present in group α are from U.S.A isolates. However, it should be noted that the variation across all these groups and subgroups is relatively small with a minimum BLAST identity of 92% when a match is present, and mostly preserved regulatory regions. Gene colour legend in **Figure 3.2**.

3.4 Discussion

3.4.1 Public Health Implications

As previously stated current WGS fine typing method relies on the core genome (Dallman et al. 2018; Janowicz et al. 2018), therefore prophage sequences have in most part been dismissed from the equation. Certain prophages such SP3, 13, and 14 may still be included in certain schemes considering their general prevalence within the majority of our isolates with a high level of conservation. Considering the importance of Stx in the pathogenicity of EHEC O157, it has also been included within PH analyses (Dallman et al. 2015; Holmes et al. 2018). However, the fact remains that, overall, due to their shared content and diversity which make prophages hard to assemble using short-read sequencing technologies, prophages have not been a key instrument in PH analyses. However, working with PHE we were able to conduct a preliminary assessment into the use of Stx-encoding prophage sequences in PH. In a study published by Schutz *et al* (Schutz et al. 2017) it was determined that the Stx2a-encoding prophage from isolate 122262 was highly similar to the one of isolate 155, thus reinforcing a geographical link that was being postulated. This was then further confirmed with the addition of Stx2a-encoding prophages from isolates PHE2 and PHEO26 to the cluster which had also been epidemiologically linked to the same geographical location (Tim Dallman, PHE, Personal communication). Finally, the stx2a-encoding prophage of isolate Z1836 of the same cluster had no links to that specific geographic location but to a farm in England. While this does not fit with prior observation, further analysis could be performed to determine whether the isolate may have indeed originated from the same location. Another noteworthy observation is the presence of this prophage across three different serotypes (O26, O55, and O157) which will be discussed in **Section 3.4.2**.

Furthermore, two large sub-clusters were detected within Stx2a- and Stx2c-encoding prophages linked to PT21/28. While it is true that the majority of PT21/28 samples were linked or derivatives of one another, the fact remains

that several samples were unrelated and yet exhibited the same subtype of Stx-encoding prophages (**Figure 3.5**, group γ). This, linked with the fact that PT21/28 is the leading PT of clinical cases in the UK and is typically associated with carriage of both Stx2a- and Stx2c-encoding prophages (Dallman et al. 2015), could lead to a method to detect PT based on the presence of these two prophages rather than the actual wet lab test. Furthermore, if one is to assume that the Stx-encoding prophages are responsible for the higher rate of human infections of PT21/28 it may be beneficial to detect such prophages when performing WGS analyses. Once more, the implication of this observation upon the evolution of Stx-encoding prophages will be discussed in **Section 3.4.2**.

The question remains whether there currently is a benefit in performing such an analysis considering the substantial cost of long-read sequencing for PH institutions as described in **Chapter 1**. At this time, the answer would most likely be no. These results are too preliminary and not yet validated to be incorporated within any PH workflow. However, with time, as long-read sequencing techniques decrease in price, and more isolates are sequenced with their prophage contents, it will become clearer whether there is indeed a strong correlation between isolates, prophage content, and certain phenotypes, ST, virulence levels, and PT. Nonetheless, this preliminary analysis seems to hint to the importance of identifying prophage content to supplement outbreak detection and tracing, as well as potential virulence genotyping. Also, while this chapter has mainly focused on Stx-encoding prophages, **Chapter 4** will take alternative approach to study the larger effect of the variety of prophages identified within EHEC O157, and how this information could also help further the understanding of EHEC O157 pathogenicity and mechanism of evolution.

3.4.2 Theory of Shiga Toxin Encoding Prophage Evolution

Due to the importance of Stx regarding the pathogenicity of EHEC, many papers have studied Stx-encoding prophages and their integration within

EHEC genomes of different serotypes (Laing et al. 2012; Xu et al. 2012; Ashton et al. 2015; Ogura et al. 2015; Holmes et al. 2018). One of the current hypotheses investigate the possibility of an EHEC O55 strain acquiring Stx2c-encoding prophages and then eventually evolving into an EHEC O157 strain, of which certain strains in turn acquired Stx-2a encoding prophages (Dallman et al. 2015). This hypothesis follows the model that prophage variation between strains is due to the integration of different prophages through horizontal gene transfer rather than prophage recombination. The data presented within this chapter seems to confirm this as clonal isolates or isolates from similar countries of origin tend to sub-cluster together. This would also suggest that the Stx2a-encoding prophage discussed previously (found in isolates 155, PHE2, 122262, PHEO26, and Z1836) is more likely a recent acquisition as it does not fit this evolutionary model. However **Chapter 4** will discuss how horizontal gene transfer may not be the only reason for large prophage variations within isolates.

Looking closer, IS elements still appear to play a large role in smaller prophage variations, and as described in **Chapter 2**, these may have large phenotypic effect. However, considering the high conservation of IS elements within specific areas of prophages, there appears to be no further proof that these are “fixing” prophages, or causing any large prophage rearrangements as previously hypothesized.

Two final points of notes: Stx cassettes, and prophage background. The data presented in this chapter allows for the hypothesis of Stx cassettes integrating within different prophage backgrounds. This is visible due to the conserved similarity of a few genes surrounding Stx genes when similar Stx subtypes are exhibited within different backgrounds (**Figures 3.3** and **3.4** show examples of such). However, a clear example of a potential Stx cassette can be observed in five Stx2a-encoding prophages, the β sub-cluster (**Figure 3.5**). In this case, these prophages exhibit a much higher similarity to Stx2c-encoding prophages rather than Stx2a-encoding ones (to which they appear to only share the cassette). Considering that the genes

close to the *stx* gene tend to be regulatory genes (Asadulghani et al. 2009), this leads to the question of whether or not entire cassettes can integrate within different prophage backgrounds. Work by Ogura *et al* (Ogura et al. 2015) suggest that the replication-related genes of Stx2a-encoding prophages influence the production of toxin. Therefore, determining the prophage background and which variant of the Stx2a-encoding prophage an isolate has may yield information regarding the potential virulence of the isolate. The concept of the Stx cassette will be discussed further in **Chapter 5**. Due to the observational nature of this study, the lack of phenotypic and metadata, and a bias in sampling, additional work is required to investigate this further.

3.4.3 Further Work

First and foremost, the key limitation of this work remains the number and diversity of sequences as well as the available metadata. Many more studies are now using long-read sequencing (Chin et al. 2013; Mikheyev and Tin 2014; González-Escalona et al. 2019) and it is becoming the norm for open access journals to require the sequence data to be made publicly available, therefore the issue of number and diversity of sequences may be getting addressed. Conversely, metadata may remain a bottle neck considering the unlikeliness of PH to start using long-read sequencing routinely. A lot of data was generated in this study (prophage similarity tree, full prophage grouping and clustering, and full gene content homology results can be found in the **Appendix II-B**), however, due to time a focus had to be made on Stx-encoding prophages. Therefore, further work could be conducted on providing similar analyses for other prophages, especially those that appear across most strains.

Finally, a phenotypic study needs to be conducted on the effects of prophage backgrounds and their variation. While knowledge of lambda bacteriophage biology can be used to understand the general behavior of Stx-encoding prophages, sequence, regulatory, and structural differences can most likely

greatly affect Stx expression, production and release. The design of an experiment to explore this is complicated. One could imagine an experiment similar to the Transposon Directed Insertion Site Sequencing (TraDIS) protocol (Barquist et al. 2016), where every gene in the prophage is disrupted in turn, and isolates are phenotyped. However, these tend to be easily observable *in vitro* phenotype tests. Instead of phenotyping one could get them RNA-sequenced which would exhibit the difference in gene and protein expression for each gene disrupted, but the cost of such an experiment would be inhibitory, including the time it would take to generate isolates disrupted in each individual gene, even if only those within Stx-encoding prophages. Therefore, a more realistic approach with current methods could involve integrating the genes observed to be part of the putative Stx2a cassette in a Stx2c background and observing the phenotypic toxicity differences. The hypothesised result being observing a toxicity pattern similar to the Stx2a β sub-cluster.

It is quite clear at this stage that much work remains to be done to truly investigate the prophage content of EHEC O157, and even something much smaller such as the background or even IS content of Stx-encoding prophages. Nonetheless, further understanding of such matters could lead to much more than just insight into EHEC O157.

4 Genome Variation Mediated Through Whole Genome Large Chromosomal Rearrangements and their Potential Effect on Phenotype

4.1 Introduction

4.1.1 Origin / Terminus of Replication

EHEC O157:H7 contains, as most bacteria do, a single circular DNA molecule with a single origin of replication (Kaguni 2011). The origin of replication typically denoted by gene *oriC* in *E. coli* is the location recognized by the initiator protein to start DNA replication (Kaguni 2011). The initiator protein then recruits other proteins such as DNA helicase and DNA polymerase to start replicating the circular chromosome as two “replication forks” extending in clockwise and anti-clockwise directions. Eventually this replicates the whole chromosome as two replichores which reach the terminus region where both forks of the replicating DNA meet (Kaguni 2011). This region is usually opposite the origin on the circular chromosome map and is determined by the terminus genes *TerA and TerB* (Neylon et al. 2005). The two chromosomes are then separated and split into two regions of the bacteria to allow division. Under rapid replication conditions, new replication forks can be started before others are completed, adding more organisational complexity to this fundamental process (Kaguni 2011).

However, this process is not perfect, and errors in the replicated DNA may occur (Fijalkowska, Schaaper, and Jonczyk 2012). It has been observed that genes closer to the origin of replication tend to not only be more highly conserved but also more highly expressed (Sharp et al. 1989; Rocha 2004). If these are key genes for the bacteria, it would make evolutionary sense for these genes to be in locations where there is the least chance for mutation occurs. Based on this line of thought it would, therefore, also be logical to assume that genes nearer the terminus may be more likely to acquire mutations including those based on rearrangements.

In EHEC O157 this leads to an interesting question regarding prophage distribution across the genome and their insertion sites. Stx-encoding prophages have well characterised insert sites (*wrba, yehV, sbcB, yecE, argW,* and *Z2577*) (Dallman et al. 2015). As stated previously, other prophages have not been studied to a similar extent, but with the advent of long-read sequencing it is now a relatively straightforward task to examine their location and content. It is apparent that the content of these prophages and where they insert could have large effects on the bacterium. As evident in the work in this chapter, the insertion sites are typically found to be closer to the terminus, potentially as they are less likely to disrupt organisation of critical genes nearer the origin. Furthermore, in this result chapter, I investigate the potential of these prophage sequences to generate chromosomal rearrangements, and why this could further select for insert sites of prophages closer to the terminus. However, prior to this, one needs to further understand how chromosomal rearrangements typically occur.

4.1.2 Large Chromosomal Rearrangements (LCRs)

Large chromosomal rearrangements (LCRs) have been extensively studied amongst species such as *Yersinia pestis,* and *E. coli* (Darling, Miklós, and Ragan 2008; Ooka et al. 2009; Darling, Mau, and Perna 2010; Raeside et al. 2014; Lee et al. 2016). The main types of LCRs are: duplications, deletions, inversions, and translocations (Raeside et al. 2014). These events are key in the evolution of any species as they can result in an increase or decrease of strain fitness by shortening the DNA, changing protein expression, or even cause replichore imbalance: where the terminus of replication is off centre compared to the origin of replication causing issues during DNA replication (Matthews and Maloy 2010; Raeside et al. 2014). One key paper (Raeside et al. 2014) focused on these LCRs in *E. coli* over a period of 25 years *in vitro*. The authors found that LCRs were frequent across their samples, especially at the start of the experiment and that fitness changes could be observed. Secondly, they observed that these LCRs, if they occurred in succession,

typically occurred in a similar region of the chromosome, and finally that the majority of those were mediated by IS element homology. They also discussed how some of the deletion events removed prophage remnants from some of the strains (Raeside et al. 2014). Another study (Ooka et al. 2009) investigated smaller rearrangements and linked those events to IS elements within prophage regions.

This chapter will investigate similar concepts but at a larger scale and from a different angle. The original papers (Ooka et al. 2009; Raeside et al. 2014) only looked at eight clinical isolates and 12 strains / populations *in vitro* respectively. This study will look at the 69 isolates previously studied, which have differing levels of relatedness such as being: from the same strain following passaging through an animal, isolated from the same farm, part of the same PT subclusters, identical PFGE pattern profiles, or simply part of the broader EHEC O157 clonal group (**Table 3.1**). Also, the previous works (Ooka et al. 2009; Raeside et al. 2014) relied on optical mapping and older short-read sequencing platforms to achieve their observations. While my current work has also utilised optical mapping to confirm certain observations, the use of long-read sequencing allows for a more precise determination of LCR boundaries and better tracking of IS elements, which, as the paper notes (Raeside et al. 2014), could not be tracked by optical mapping. Finally, EHEC O157 has on average a larger genome size, when compared across the *E. coli* species, due to its higher prophage content, and as noted (Raeside et al. 2014) shows a lesser degree of genome size reduction over time when compared to other *E. coli* serotypes. This indicates that EHEC O157 is possibly selecting the higher fitness cost of a larger genome in favour of maintaining its prophage population. As such, modularity within the prophage content of an isolate without requiring additional genome additions would be highly beneficial to the bacteria.

4.1.3 Optical Mapping

Optical mapping involves the partial single stranded digestion of genomic DNA with a restriction enzyme and then end labelling with a fluorescent probe (Ravindran and Gupta 2015). The resulting labelled fragments can then be analysed by florescence microscopy or by being pulled through a detection pore. The fluorescent points generate an optical map of the restriction sites for each DNA molecule. These maps are then overlapped based on fragment sizes to determine the consensus optical map (Ravindran and Gupta 2015). Even in the age of long-read sequencing, some advantages of optical mapping remain, and these include the ability to identify structural changes, especially duplications, which are larger than most long sequencing reads can span. Optical mapping can analyse tens of thousands of single genomes in a single run. Unfortunately, this advantage is offset by the fact it is technically difficult and expensive. The development of “ultra”-long reads on the MinIon platform (some recorded as long as 1 Mbp) (Oxford Nanopore 2017) may provide an alternative way to analyse LCRs. However, at present, our own attempts at replicating such “ultra”-long reads for a few of our samples were unsuccessful. Therefore, optical mapping was used on specific samples to confirm some of the observations made from the sequencing and in one case, detected a large duplication for a genome that PacBio sequencing was unable to close with high confidence due to the presence of this duplication.

4.2 Methods

Chapter 4 takes over the data from **Chapters 2** and **3** therefore the method for prophage calling is the same as before, however, only 69 EHEC O157 sequences were used in this analysis as the rest were either sequences of Stx-encoding prophages or unclosed genomes.

All code written for the purpose of this chapter can be found at: https://github.com/SharifShaaban/PhD-Code (subfolders Appendix III-A, III-B, and III-C). The following tool versions were used: EMBOSS 6.6.0, Perl 5.18.1, BLAST+ 2.4.0, R 3.2.2, Python 2.6.6, Easyfig CL 2.1, and Circos 0.69.

4.2.1 Whole Genome Alignments

Using the same method introduced in **Chapter 2** whole genome alignments were generated but similarly to **Chapter 3** the alignment shows homology between genome areas. Once more these were carried out using EasyFig (Sullivan, Petty, and Beatson 2011) but the BLAST (blast + version 2.7.1) (Altschul et al. 1990; Camacho et al. 2009) matches minimum length threshold was raised to 20000 bp. This was done to avoid background noise in the figure and make it more readable (if the threshold was set lower, identical IS elements and other smaller areas of homology across the chromosome would have been highlighted thus making the figure too hard to read). All prophages were given a neutral beige colour, with Stx-encoding prophages given a colour representing their Stx subtype (red for Stx-2a, green for Stx-1a, blue for Stx-2c, and cyan or Stx-2d). Scripts to generate genbank files with these features can be seen in **Appendix III-A**. The order of genomes in the figures was determined through a manual iterative process where the alignment was run, similar genomes were ordered adjacent to one another, and the alignment run once more, until genomes were sorted alongside similar prophages whilst highlighting inversions. In the case of

Figure 4.2 isolate 9000 was depicted twice in order to be more easily compared across all the sequences.

Four different alignments were produced. The first one regroups all 69 whole genome sequences, and serves as an overview of the LCRs as well as a rapid comparison of any potential arrangements specific to a certain chromosomal region or group of isolates.

The second alignment is between eight isolates that were originally typed using PFGE EHEC O157 typing as described in **Chapter 1**. Their PFGE types can be found under the isolate names in **Figure 3.2**. These isolates were obtained during the IPRAVE project, a Wellcome Trust-funded epidemiology programme in which there was a Scotland-wide collection of *E. coli* O157 isolates from cattle, all were typed and a subset subsequently sequenced. At the time of sampling (2002-3), one PFGE pattern, profile C (updated to be labelled A 11b), of PT21/28 strains was common across cattle and human *E. coli* O157 isolates (IPRAVE members, Personal communication) (Chase-Topping et al. 2008). Anecdotally (Dr Lesley Allison, SERL, Personal communication) *E. coli* O157 PFGE profile types were stable in humans (with three main types: A, B & C) until ~2005 when there was an expansion of types (>30), although profile C remained abundant. Further characterisation of isolates using “short-read” sequencing, and the SNP-typing method (using SnapperDB (Dallman et al. 2018)) described in **Chapter 1** (data not shown) has provided evidence that isolates with different PFGE profiles can be closely related at a SNP level. For example, isolates Z910 and Z563 actually have an identical SNP profile (based on the typing methodology described in **Chapter 1**) but different PFGE profiles. The other isolates in this alignment were selected to be PT21/28 yet include a variety of different PFGE profiles, while still exhibiting a certain level of relationship using SNP typing (majority of isolates are within the same 25 SNPs single linkage cluster, which has been shown to represent isolates from the same outbreak in rare cases, while all isolates are within the same 50 SNPs single linkage cluster). The alignment of 12 isolates, has isolate 9000 represented

twice, this is to mark all the XbaI restriction enzyme sites across the genome. These sites define the PFGE profile as the restriction enzyme in the laboratory's methodology cut at these specific sites (CDC Pulsenet Standard Operating Procedure manual). Representing isolate 9000 twice also make it easier to compare profile C (A 11b) with the other PFGE profiles. Inkscape (Inkscape 2019) had to be used alongside EasyFig to superimpose the XbaI restriction enzymes sites over the predicted prophages (no other edits where conducted on this figure). Also, the BLAST hit representing the XbaI restriction sites (six base pairs long) had to be expanded to 5000 bp to be visible on the figure (2500 bp left and right of the XbaI sites were added to the coordinates).

The next two figures show the alignment of sequences from isolates with known relations: the "X" animal trials isolates and the "Y" outbreak isolates respectively.

The "X" animal trials isolates were from two experiments performed by the EHEC research consortium (including the Moredun Research Institute and the University of Edinburgh) in order to investigate "super-shedding" (Fitzgerald et al. 2019) and the importance of Stx2a for transmission of a "super-shedder" strain. This was conducted using isolate 9000 as the inoculum isolate in the first trial, and a version of isolate 9000 with a repaired Stx2a-encoding prophage (IS element removed from the Stx2a gene, **Chapter 2**) in the second.

The "Y" outbreak isolates were isolates 180 and 644. WGS analysis of this outbreak was fully analysed in (Cowley et al. 2016). The authors of that paper found that outbreak "Y" was two separate clinical events eight weeks apart. They discovered through "short-read" WGS that the strains for each separate event were only three SNPs apart. However, they exhibited two different phage type profiles (PT8 for isolate 644, the earlier isolate, and PT54 for isolate 180, isolated eight weeks later).

The reason behind choosing these two sets of isolates for further alignments and study is, as stated previously, the knowledge that these isolates are

directly related, in one case must originate from one another (cattle colonisation) and in the other have been demonstrated to share a very recent common ancestor (outbreak). Therefore, any LCRs observed between these strains allow for greater insight in LCR occurrence during isolate evolution.

4.2.2 *In silico* PFGE

Considering our usage of dated PFGE data in our selection of isolates for this analysis as well as the hypothesis proposed in this chapter, it was deemed important to further understand the PFGE profiles and the differences seen in profiles between isolates. As such *in silico* restriction digestion was carried out using the Sequence Manipulation Suite V2 (SMS) Restriction Map tool (Stothard 2000). The digestion sites for XbaI were extracted, and putative fragment sizes were determined. Any fragments below 48500 bp were dismissed as these would have been hard to differentiate and likely ignored in an actual PFGE analysis (Lesley Allison, SERL, personal communication). The remaining fragment sizes were used to create a dot plot (column scatter plot) with a violin outlay in R (R Core Team 2019) using the ggplot2 package (Wickham 2009). Code is available within **Appendix III-C**. Furthermore, the figure was edited in Inkscape to remove the default legend, and sharpen labels for clarity. This visualisation method, while unorthodox for this type of data, was chosen for a couple of reasons. First, all available tools which did *in silico* PFGE with simulated gels visualisation had limited sequence input sizes, and therefore, could not be used appropriately. Secondly, this visualisation mimics a gel to a certain degree. “Fragments” are separated by size in a vertical fashion, creating a similar profile to that of a PFGE. A key difference is the approximations made for this diagram. For one, the graphics package attempts to group “fragments” for visualisation purposes, and therefore bins together different “fragment” sizes causing the figure to not be as exact as an actual gel. However, it is suitable for this specific purpose. Sample order was determined alphanumerically based on isolate name.

4.2.3 Chord Diagram / Circos

As will be discussed, further whole genome alignments of the 69 isolates demonstrated LCRs, even between closely related strains (through SNP-typing). Therefore, a method to visualise areas of homology was required (since as stated previously areas of homology typically mediate LCRs). A successful visualisation tool was Circos (Krzywinski et al. 2009), a chord diagram generator. Typically, chord diagrams utilise an outer circle with relationship between data points (given within a matrix) drawn as arcs connecting the different areas of the outer circle.

A script was written to generate the Circos input file (scripts and example input files in **Appendix III-B**). While it contains some formatting content, most of it is the data matrix. This matrix was designed so that the outer circle was divided at prophage boundaries, with prophages being coloured red and the rest of the chromosome in blue. Prophages with predicted overlapping boundaries were merged into single red blocks but can be identified by comparing the Circos plots to the whole genome alignments where they will appear as overlapping blocks. The BLAST score of querying the genome sequence to itself was used to inform which Circos segments are linked and how large that linkage is (representing homology). Any BLAST hit that exhibited 98% homology and that was larger than 5000 bp was included into the matrix, the length of the match representing the width of the segment linking outer circle areas. Furthermore, only hits between the following coordinates were recorded: 1500000– 4500000, and these had to be within prophage areas. These thresholds were determined based on the following rationale:

- If hits of any size were allowed, too many hits were recorded across the whole genome causing a lot of noise. Also, hits larger than 5000 bp accounted for the larger LCRs observed, and of interest. This would disagree with the paper previously mentioned and its focus on IS elements mediating LCRs. However, this does not mean that it was not the IS elements within the denotated areas of homology that

mediated the LCRs (more in the **Discussion** section). Also, unlike **Section 4.2.1** the aim of these figures was not to simply show LCRs but point to the homology that could potentially be mediating them, thus the need for a smaller threshold than previously.

- Homology hits found in the areas outside the given coordinates were deemed unlikely due to their potential effects on genes close to the origin of replication, which as stated previously could have detrimental effect on the bacterium.
- Hits outside of prophage areas, while potentially relevant were dismissed due to the focus of this study on prophages (however, figures showing hits within non prophages areas can be found in **Appendix III-B**).
- A BLAST identity score of 98% was chosen. This was chosen partially due to the aforementioned paper finding LCRs mediated by genes with 96% homologies and that other sources have documented recombination requiring DNA areas of 100% homology. However, these sources discuss DNA areas as small as 100 bp. Therefore, assuming a large area of homology, that score of 100% can be lowered as one can assume that within that large area of 98% homology, there will be smaller areas of 100% homology.

An example figure removing the homology length and its location thresholds can be found in **Appendix III-B** to better illustrate these points. Furthermore, it should be noted that prophages could have areas of homologies to themselves, but this inaccurately doubled the width of the connecting segment. Finally, Circos does not allow for homology hits to represent the area of the prophage which they match to, the connecting segments simply originate and end at the earliest available location within the matching outer circle area.

4.3 Results

4.3.1 Complete Whole Genome Alignment

The whole genome alignment of 69 isolates, from different geographical locations, source, temporal periods, and hosts, of different phage types, PFGE profiles, and phenotypes was generated and is presented in **Figure 4.1**. Even with the restrictive thresholds used to diminish noise in the figure, it remains a figure with a wealth of information that can be hard to read. Once more due to time limitations, it was not possible to explore all the insight offered even through such a simple alignment. One observation of note is the high level of identity across most of the genomes regardless of the isolate, even with a minimum of 98% identity. Three main sources of variations can be observed: prophage presence and absence, LCRs at, or near, prophage boundaries, and IS elements (identifiable by small gaps within large homology blocks). Most of this variation occurs near the terminus of replication (which would be around the middle of the alignment as the starting point is the origin of replication). Many isolates (starting from isolate 9000 to isolate Z1813) show even less variation. However, these are all PT 21/28, showing a selection bias in the isolate pool that will be addressed in the **Discussion**.

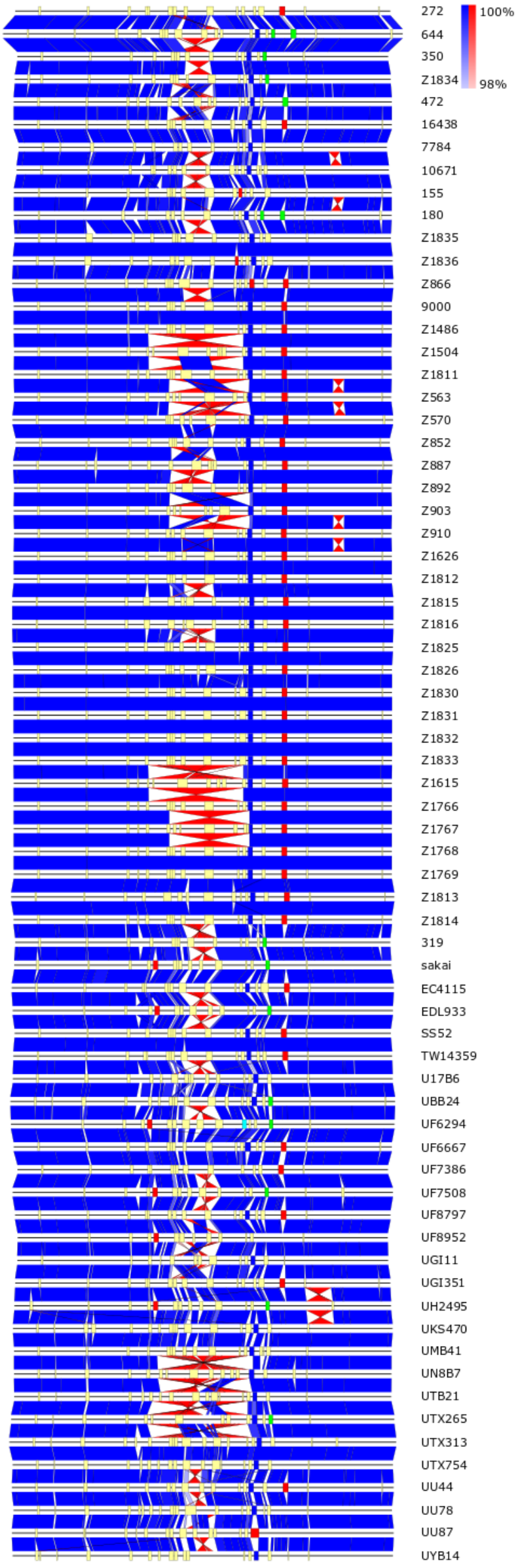


Figure 4.1 Whole genome alignments of 69 EHEC O157 isolates. Genomes are designated with a black line, with prophages marked as blocks on these lines. Stx-encoding prophages are marked in red for stx2a, blue for stx2c, green for stx1a, and cyan for stx2d (only found in UF6294). All other prophages are beige. Between the lines in blue or red are BLAST matches, with the shade of colour determining the level of match ranging from 98% to 100% identity, with red indicating inverted matches. A few main observations can be made from this figure:

- There is variation within the prophage population shown by prophage presence and absence between isolates, with little apparent geographical concordance (seen by the wide diversity within US isolates, prefixed by the letter U, however, due to the lack of provenance of these isolates this needs much further investigating).
- Most LCRs are in the form of inversions and occur near the termination of replication. Even certain cases of what appear to be translocation can technically be explained by multiple inversions (see **Section 4.4.2**).
- The Stx-encoding prophages appear to have well maintained insert sites, especially Stx2c-encoding prophages (in blue) which when present are always within the same insert site. Furthermore, these Stx2c-encoding prophages appear to be, at times, involved in LCR boundaries, unlike Stx2a-encoding prophages (in red) for which the common insert site is further from the terminus or replication.
- Phage type 21/18 isolates appear to exhibit less variation within it than all other isolates. However, this could be due to the large selection bias in the isolates and therefore, requires much further investigation with many more isolates of different provenance and with known relationships.

4.3.2 Related Whole Genome Alignments

The first of the smaller alignments (**Figure 4.2**) demonstrates that LCRs can affect PFGE profile due to the high number of XbaI restriction enzyme sites across the whole genome, and that the majority of LCRs between these quasi-related isolates is still found near the terminus of replication. Furthermore, one can observe that certain smaller changes that do not alter genomic size or alter the arrangement of the XbaI sites, such as those between isolates 9000 and Z1486, do not cause a change in PFGE, which is as predicted. However, smaller genomic differences that are not LCRs and do not meet the previously stated limitations, can still cause a different PFGE pattern to occur (e.g. isolates Z1486 and Z570). Furthermore, when looking at isolates Z563 and Z910, an LCR is the only cause for a different PFGE profile considering they exhibited no difference at a core SNP level and shows very little other accessory genome differences based on the alignment.

Furthermore, the *in silico* PFGE data appears to confirm this. When comparing samples 9000 and Z1811 (which are shown to be closely related when using SNP-typing) it is apparent that this difference in PFGE profile is mainly due to the LCR. Firstly, the rest of the genome looks highly similar with the only other noticeable differences being potential IS element movement. They differ by a simple large inversion. When looking at the fragment size distribution (**Figure 4.3**), only two fragments differ between the two isolates (one for each site at the edge of the LCR). And finally, when investigating the coordinates of the restriction sites (**Appendix III-C**), the two “fragments” which differ between the two isolates are the ones using the sites at the edges of the LCR. The same can be observed when comparing isolates Z563 and Z910. Notably, in both these cases, “fragment” sizes normally shift by small numbers (<50 bp) or sizes that seem to indicate IS elements (~1200 bp).

This, however, is not the complete story. When comparing isolates 9000 and Z910, the main difference appears to be a relatively small LCR and a few areas of divergence, the difference in PFGE profile does not appear at the coordinates near the LCR (**Appendices III-B** and **III-C**). Furthermore, this LCR appears to

have no effect on PFGE “fragment” sizes when looking at the coordinates of the XbaI sites which are near the LCR. Therefore, in certain cases, minor changes as observed between those two isolates can still cause PFGE differences. This indicates that the overall picture probably is a combination of both LCRs and small chromosomal differences.

Figure 4.4 of whole genome alignments pertains to isolates with known relationships and therefore more insight can be gained in terms of the type of re-arrangements that are occurring routinely. Due to the closeness of these isolates, as previously described, all homologies found that were larger than 20000 bp, as shown in the figure, are of 99% identity or more.

First (**Figure 4.4 A)** is the alignment between isolates 644 and 180, which confirms all that was discussed in the previously mentioned paper (Cowley et al. 2016). There are four sources of variations (not including SNPs). These are (1) second prophage of isolate 644 being absent in isolate 180, (2) a large inverted duplication, (3) a part of one of the prophages within the boundaries of the duplication present in isolate 644 missing in isolate 180, and (4) the eighth prophage of isolate 180 missing from isolate 644.

The second alignment (**Figure 4.4 B**) in the figure shows an even higher level of homology which causes the figure legend and colour legend to be ambiguous, as all homologies larger than 20000 bp found were closer to 100% identity than 99%. Therefore, while some SNPs may exist between the isolates, these are few, and the main source of variation is LCRs in the form of inversions, one spanning nearly 1.2 Mbp (between isolate Z1615 and isolate 9000). In both cases, all the LCRs observed are near the terminus of replication.

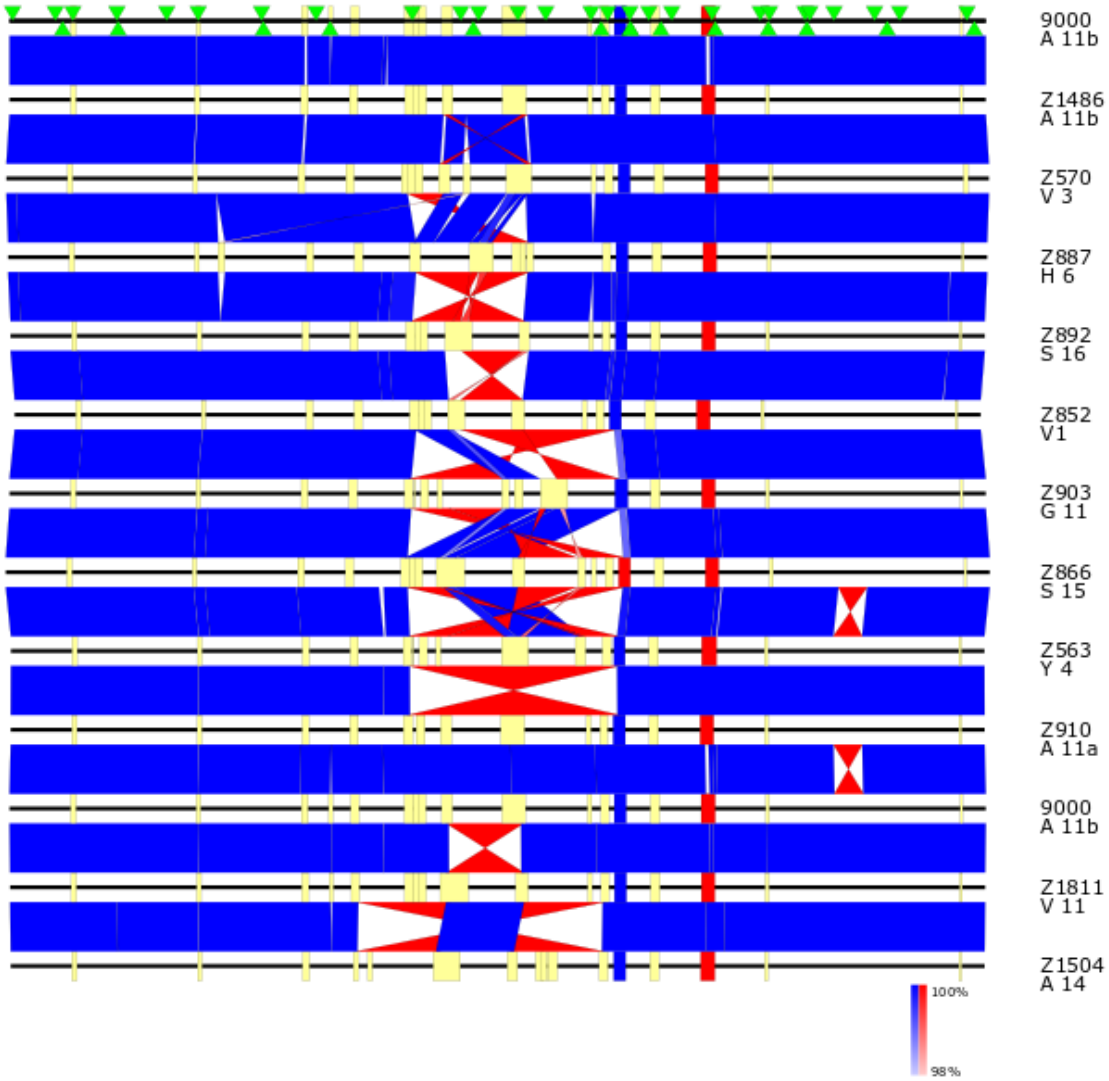


Figure 4.2 Whole genome alignment of twelve isolates, with isolate 9000 represented twice for visualization purposes. In green are markers indicating the XbaI restriction enzyme cutting sites. Coloured blocks along the black lines representing the chromosomes, are prophages with Stx2c- and Stx2a-encoding prophages being blue and red respectively, with all other prophages being beige. BLAST matches are represented by shades of blue links between the chromosomes (minimum 98% identity), and red links for inverted matches. Beneath the isolate names are their PFGE profile as defined by SERL at the time. Similarly to prior alignments, the main source of variation appears to be LCRs in the area of the terminus of replication. While not relevant to PFGE, the alignment between the final two isolates (Z1811 and Z1504) is slightly misleading due to the way the inverted homology hit is hidden behind the blue homology hit. It is worth noting that this is actually three separate matches, and not a case where a palindrome exists within a LCR.

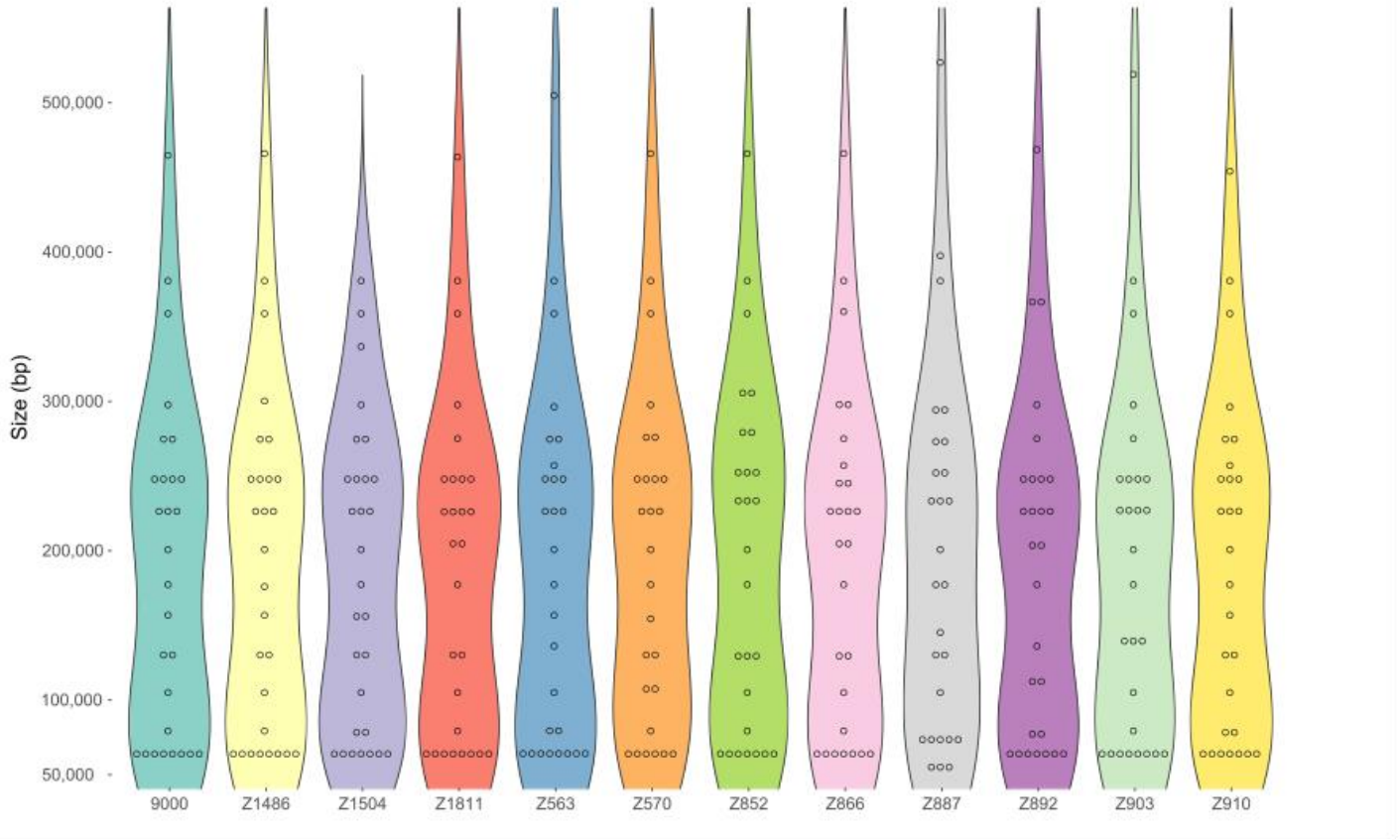


Figure 4.3 Dot plot (column scatter plot) of *in silico* predicted PFGE "fragments" using the XbaI restriction enzyme. Dots represent predicted fragments and their size if PFGE profile conducted for isolate. Isolates with the same dot profile are thought to share a *in silico* PFGE profile, such as isolate 9000 and Z1486.

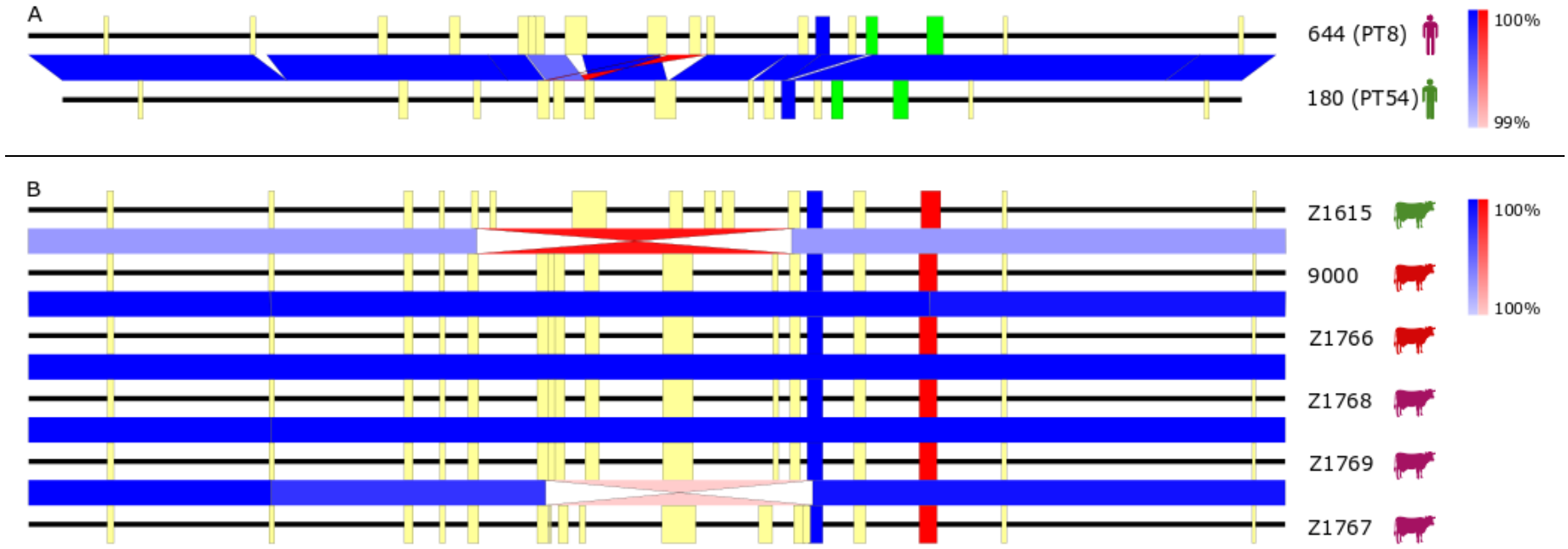


Figure 4.4 A. Whole genome alignment of isolates 644 and 180. Matches are represented in different shades of blue ranging from 99% to 100% identity, and in red are inverted matches. Prophages are represented by block, with Stx-encoding prophage being blue for Stx2c and green for Stx1a. The main differences between these two isolates are the second prophage of isolate 644 missing from isolate 180, a large section of the eighth prophage from 644 missing in 180, an inverted duplication containing prophages and non-prophage genomes in 644 not present in 180, and the eighth prophage from isolate 180 missing from isolate 644.
B. Whole genome alignment of isolates 9000 (inoculum isolate in the first trial), Z1766 (inoculum isolate in the second trial, identical to isolate 9000 but with a repaired Stx2a encoding prophage (coloured in red), Z1615 (isolate obtained from the first trial), and Z1767, Z1768, and Z1767 (isolates obtained from the second trial). Due to the high homology levels, the colour legend indicates in lighter shades hits that are still closer to 100% homology than 99%. As such the main differences between those isolates are large inversions with their boundaries within or near prophage areas present in isolates Z1615 and Z1767.

4.3.3 LCR Homology Determination

The previous figures identified LCRs around the terminus of replication. As previously stated, one mechanism for generation of these are areas of homology that may allow homology-based recombination, for example behaving as large inverted repeats. As such the following two figures (**Figure 4.5** and **4.6**) investigate the homology present between the isolates aligned in **Figure 4.4**.

The first figure is divided into two parts and compares the homology present in isolate 644 and isolate 180. As previously established the key differences between these two isolates were few, yet looking at the homology presence it becomes apparent that these changes have a large effect on the amount of homology present within each isolate. Isolate 644 has more areas of homology (which is logical considering the presence of an inverted duplication of about 200 kbp within it), thus having a potential effect on the amount of potential recombination that can occur. However, even though isolate 180 has fewer areas of homology, considering that these are at least 5000 bp long (full length of homologies can be found in **Appendix III-B**) it still has a high potential for recombination. Interestingly, assuming that multiple different inversions can occur, this can lead to a different pattern of LCRs. For example, if a large inversion occurs, and is then followed by another inversion involving one of the areas of homologies that was part of the previous inversion, and another area of homology that previously had not inverted, this would result in a isolate with an apparent translocation (such as the one seen in Figure 5.1 for isolates Z1834 and 472). Therefore, due to the number of large areas of homology these could lead to complex LCRs simply through multiple inversions, which as we now can observe, appear to be the most prevalent type of LCR within the studied population of isolates.

The second figure is more straightforward as the LCRs observed are simple inversions, for which the homologies labeled A and B in **Figure 4.4** can account for given that an inversion occurs between these areas of homology. The potential meaning of these inversions and the effect they could have will be further discussed in the next chapter.

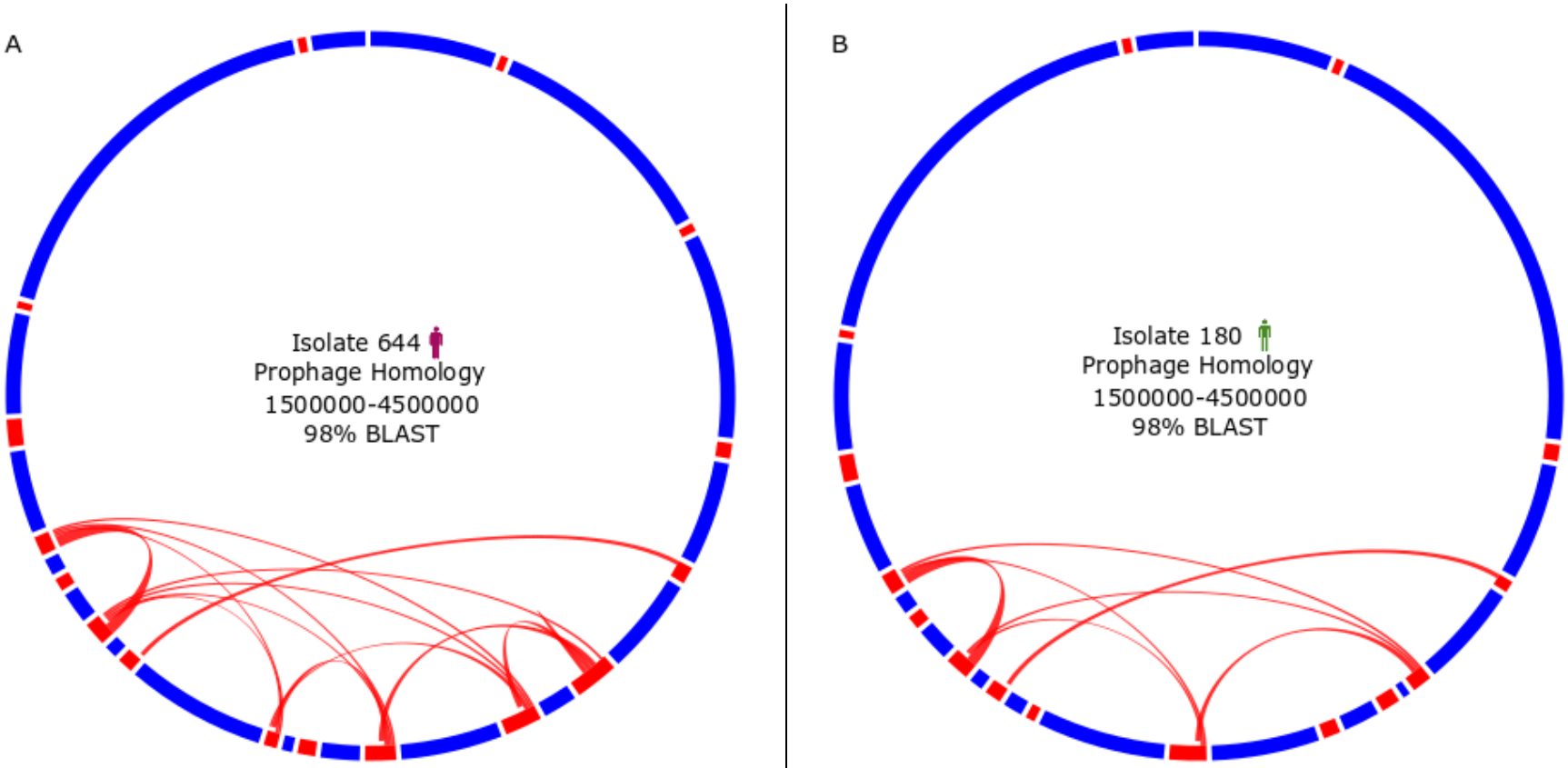


Figure 4.5 A. Circos plot showing homology between prophage regions (shown in red blocks) using red links between blocks for isolate 644. Homology had to be present between the coordinates marked in the legend, and at least have a 98% BLAST identity match.

B. Same as above but for isolate 180. This diagram shows a lesser number of sites of homology between prophages which is due to the lack of the duplication region observed in **Figure 4.4 A**.

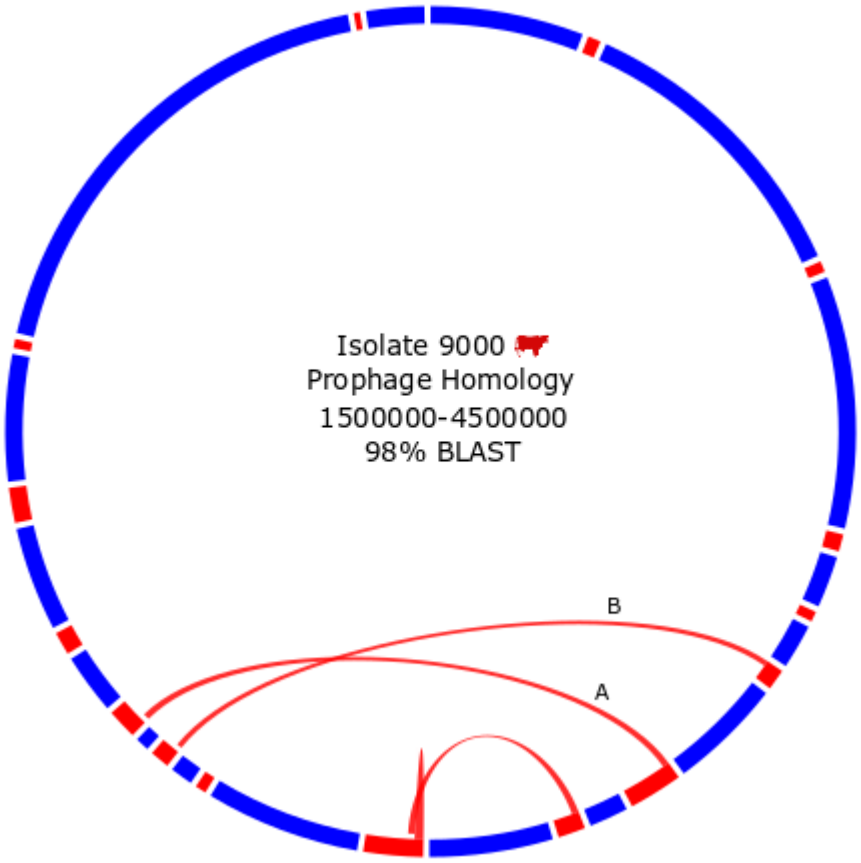


Figure 4.6 Circos plot showing homology between prophage regions (shown in red blocks) using red links between blocks for isolate 9000. Homology had to be present between the coordinates marked in the legend, and at least have a 98% BLAST identity match. Marked by letters are homologies that may mediate the inversions observed in **Figure 5.4 B**, A being the homology that could potentially mediate the inversion observed between isolate Z1766 and Z1767, and marked B is the homology potentially mediating the inversion observed between isolates 9000 and Z1615.

4.4 Discussion

4.4.1 EHEC O157 Prophage Diversity

In the literature the main paradigm for prophage diversity in EHEC O157:H7 is the integration or excision of prophages from the bacterial chromosome (Asadulghani et al. 2009). While, this work supports this hypothesis to a certain degree, the genome assemblies made possible by long-read sequencing point to another important level of variation: Large Chromosomal Rearrangements (LCRs). Furthermore, these mainly had their boundaries within prophage regions. As such when an LCR occurs it has the potential to reshuffle the prophage content. For example, looking at **Figure 4.4 B**, at the inversion between isolate 9000 and isolate Z1615 one can clearly see that the prophages at the inversion boundaries are of different sizes (11th prophage from the left for isolate Z1615 and 12th prophage from the left for isolate 9000). Therefore, even though the areas of homology mediate the inversion, the sequence preceding and following it offers different "configurations" for these prophages. Hence, the prophage content can differ and may result in variation in regulation or capacity to mobilize prophage content; all this without the integration of new prophages. This could have a large impact on the phenotype of isolates, especially considering that such an inversion occurs within a Stx2c-encoding prophage (**Figure 4.4 B**, inversion between isolate Z1766 and Z1767).

Secondly, as previously established, IS elements can have an impact on prophages and phenotypes (Manuscript submitted to Nature Microbiology) (Ooka et al. 2009), offering another method for affecting prophage activity without the integration of novel genetic content. Furthermore, IS elements have been hypothesised to be involved in LCRs in *E. coli* (Raeside et al. 2014; Lee et al. 2016). However, in our data IS elements detected within prophages (**Appendix II-B-Temp-IS BLAST Res**) were not near the boundaries of key observed LCRs such as those between isolates 9000 and Z1811, or isolates 9000 and Z1615, or isolates Z1767 and Z1769 (**Appendix III-A-BLAST Outputs Examples**). Some of the prophages involved within

certain of these LCRs did not have IS elements detected within (**Appendix II-B-Temp-IS BLAST Res**). However, it is important to note that IS detection was reliant on a database of IS sequences as described in **Chapter 2**. Therefore, certain IS elements may have been missed in this analysis. This does not mean that IS elements are not involved in LCRs, however, it does offer the possibility that IS elements are not the sole mechanism behind LCRs. However, this would require further investigation (**Section 5.2**).

Another interesting possibility to consider is how the combination of multiple inversions from different homology sites can result in rather complex LCRs that could be confused with prophage translocation, as mentioned in the **Results** section. This offers a potential explanation for certain prophage translocation events which could otherwise be considered to arise from prophage excision and integration events rather than multiple LCRs. This also leads to an important realisation about how unreliable pre-WGS typing methods were. These LCRs could easily change PFGE profiles and potentially phage type. In the examples of isolate 644 and isolate 180 such an LCR occurs yet these are within two highly related outbreaks. This then leads to the question of how these LCRs might affect current WGS typing methods such as SNP typing, and cg-MLST typing (or even traditional MLST). The answer is that they should not as the prophage regions are normally dismissed from core genome analysis, its reshuffling should have no effect on the typing method (as these would be dismissed as areas of high recombination), while the inversion of the core genome would not be detected through read mapping. Therefore, while related strains could look highly different using the techniques presented in this chapter, they would still be detected as related through other methods (as described in (Cowley et al. 2016) paper). This, however, does not mean that isolates with LCRs would have the same phenotype as will be discussed.

4.4.2 An Isolate's Potential for Recombination and its Effects

The paper studying the "Y" outbreak offers a unique phenotype switch between the two isolates. Isolates related to isolate 644 were found in fewer numbers and caused a less symptomatic outbreak, which was then followed by many isolates related to isolate 180 causing a wider outbreak with more severe symptoms (Cowley et al. 2016). As previously stated, these only had a three core SNP differences. The authors of (Cowley et al. 2016) proposes that the different phenotype is potentially due to the fitness stress a large duplication could put on an isolate.

Based on the findings of this work, while it may appear that the two related isolates must have undergone quite a change to reach the level of differences shown in **Figure 4.4 A**, I hypothesise that this could be reached through three simple inversions. As previously stated, three inversions involving one common prophage can result in a final "configuration" that looks similar to a translocation (**Figure 4.7**). A potential mechanism could involve DNA breaks at the homology sites, if the duplication seen in isolate 644 contains one of these homology sites, it could be lost during such an inversion event. Therefore, with three simple inversions, one can potentially explain most of the chromosomal differences seen between the two isolates.

These events did supposedly take up to eight weeks to occur (Cowley et al. 2016). However, based on the "X" animal trials, these LCRs can occur at a much higher speed, and based on the literature these inversions have been seen *in vitro* (Raeside et al. 2014). It is important to note that the David Gally group conducted further testing on these inverted isolates to determine whether they exhibited different phenotypes, and the results showed that bacterial fitness, type three secretion systems, and toxin expression were all influenced by the isolate's "configuration" (Manuscript submitted to Nature Microbiology). This leads to the question of how common these inversions are, the frequency at which they occur, and the survivability of the isolates.

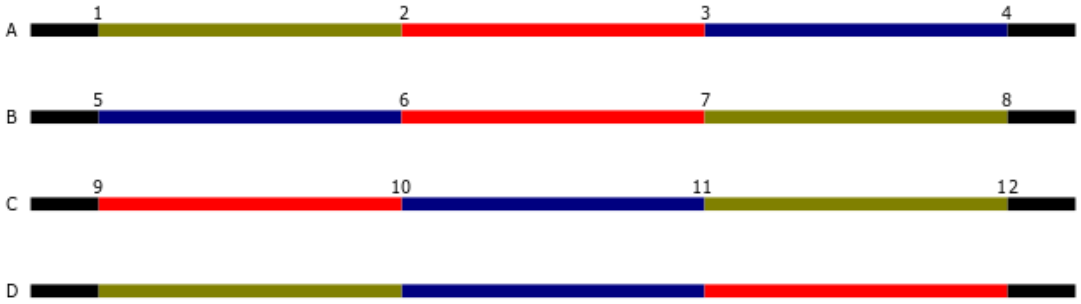


Figure 4.7 This diagram shows how the result from three inversions can appear like a translocation. In this example, every numbered intersect between a coloured region is assumed to be an identical region of homology. If the first inversion was to occur between the outer most homology regions (labelled 1 and 4), it would result in the "configuration" labelled B. If this is followed by an inversion within homology regions within the initial inversion (labelled 5 and 7), this would result in "configuration" C. A final inversion between the outer most areas of homology once more (labelled 9 and 12) would result in "configuration" D, given the impression that the green segment has translocated without being inverted, while the red and blue segment were inverted. In this example, we assume, 4 identical areas of homology, but in real world examples, similar results can be achieved by having areas of homology within the same prophages to other prophages as shown in **Figure 4.5** and **4.6**.

4.4.3 The Age of Phage

Several publications (Darling, Miklós, and Ragan 2008; Ooka et al. 2009; Raeside et al. 2014; Lee et al. 2016) have observed genome changes occurring in E. coli O157, both following cultures *in vitro* (Raeside et al. 2014) and PFGE variation after animal colonisation (Hänninen, Hakkinen, and Rautelin 1999). This makes sense considering the large number of areas of homology in these strains, especially as large areas of homologies provide a higher likelihood of inversions occurring. However, while IS elements have been hypothesised to be the main source of homology behind LCRs, this work demonstrates that most probably larger areas of homology (which could not have been determined using the traditional sequencing techniques) are more likely responsible. Considering the large number of areas of homology across the chromosome, LCRs should be common. Yet as previously stated, inversions involving areas near the origin of replication would have high odds of being negative for the organism. Following the same logic, and assuming a random selection of LCRs, some recombinatory events, even near the terminus of replication, could still be disadvantageous to the bacterium. Therefore, while one may see many LCRs occurring *in vitro*, given that these are grown in nutrient rich media with low stress, a lot of these different chromosomal “configurations” may simply not survive *in vivo*. Furthermore, one needs to consider the selective pressure applied on the bacterium *in vivo* further decreasing the amount of potential “configurations” that can be observed in real world isolates. Host jumps occur but are relatively rare. Therefore, it could be hypothesised that if no host jump was to occur, the bacterium observed is already in its optimal “configuration”, and other “configurations” seen would be those having a lower phenotypic effect, as seen in the “X” animal trials which took isolates that were obtained from cattle supplying them back to a cattle host. A potential experiment to test this theory would be an animal trial in which one would induce a host jump to another animal/host/environment than cattle and observe whether LCRs with a phenotypic effect are observed. This would agree with results seen in the literature, however, these were looking at PFGE profiles in chicken with a

different organism (Hänninen, Hakkinen, and Rautelin 1999). This could also explain the changes seen in the “Y” outbreak isolates where, hypothetically the earlier isolates might not have had much time to adapt to the new human hosts before being isolated, while further isolates might have been passed through an infected worker at the food outlet. Another potential test would be sampling of the different stages of the terminal rectal colonization, and to observe if different “configurations” could be isolated and sequenced at separate times.

This hypothesis would require a large amount of testing, such as the animal trial proposed, but other cases of related human outbreaks with a temporal difference may also be useful to investigate this. Furthermore, it is not necessarily applicable to host jumps, but such LCRs could allow for different “configurations” which are more suitable for different environments within a single host (such as either the rumen or the colon). Genome plasticity is essential for many bacteria, but particularly for *E. coli* due to the diverse environments it can live in. However, this ability of LCRs to generate the required plasticity is yet to be demonstrated.

The aforementioned theory does seem to be supported by the PFGE data observed. The different PFGE profiles observed, even between isolates that appear to have indistinguishable core genomes, and that would, therefore, have been thought to be from the same outbreak, are mainly caused by LCRs. This also raises a myriad of questions related to the phenotypic effects of LCRs. Certain PFGE have been more prevalent and may be associated with more severe clinical conditions. However, to truly investigate further, one would need to do an adequate association study between the clinical metadata and PFGE to see if two different PFGE types that are obviously caused by an LCR, have different typical clinical outcomes. This hypothesis can be further extended by assuming that specific “configurations” may be more suited to infect human or cattle hosts (such as the “configuration” yielding PFGE type C 11b), thus explaining how a PFGE could be persistent within the EHEC O157 population. Furthermore, this hypothesis still allows

other chromosomal difference to have different clinical outputs, as it was observed in **Figure 4.2** and **Figure 4.3** that smaller differences can occur and not affect the PFGE profile.

However, one must wonder whether these inversions actually occurred in the host or while the isolate was being grown for sequencing (or PFGE typing). This could be tested by single cell sequencing, which does not require cells to be cultured, however, these experiments have many caveats such as a high PCR amplification bias and a higher risk for contamination, noise, and uneven coverage, which would therefore require multiple runs which would be an expensive experiment. However, this could potentially answer another question: are the populations of cells sequenced from a single isolate homogenic? Once more, assuming LCRs are as common as this study suggests, and that some may have minimal effects on phenotype if no host jumps occur, it would be sensible to assume that populations isolated would not be homogenic. Especially considering that while higher levels of homology may not always be beneficial to a bacterium (e.g. isolate 644), it does give it a lot more versatility, and potential for rapid change. Therefore, within a given host, the observed sequence may solely be one of the "configurations" present within the host, while other "configurations" may be present in the population background, and then grow to the majority when the condition favours them. Single cell sequencing could investigate this. However, this could also be investigated through MinIon sequencing, as it offers the possibility for ultra-long reads, potentially covering LCRs in one read, so that one could detect an LCR even if it is present at a low percentage within a population. This theory allows for a myriad of hypotheses. Such as that certain strains, with a higher potential for LCRs, may make more successful multi-host isolates.

In conclusion, from the work presented here, one can observe how the prophage population found in the genome of O157 opens a wealth of potential for the bacterium to generate different "configuration" that could affect host specificity, pathogenicity, and strain fitness. However, a lot of work

remains to be conducted on these hypotheses, with many different potential experiments that could give insight into the LCR population found within EHEC O157.

5 General Discussion

Considering that all results chapters have an individual discussion section, this overall discussion will summarise some of the data observed that has not been further investigated, present general hypotheses regarding the EHEC O157 prophage content, and discuss the potential of the data presented in this thesis for application in the field of Public Health (PH).

5.1 Prophages: The Key to Genome Modularity

Multiple hypotheses were introduced throughout this work. While the next section will investigate the limitation of this work, and the future work required to investigate these hypotheses, first, this section will regroup some concepts that were only briefly touched upon. One thing that is true for the majority of hypotheses presented in this paper is that prophages are central to the ability of EHEC O157 to be a worldwide zoonotic threat (Akashi et al. 1994; Dallman et al. 2015; Cowley et al. 2016). This is mainly exhibited through their ability to introduce *stx* genes into the EHEC genome as described in other works (Ohnishi, Kurokawa, and Hayashi 2001; Asadulghani et al. 2009; Ogura et al. 2015), but also through the potential versatility prophages, in conjunction with IS elements, offer. As presented in this work Stx-encoding prophages are only part of the picture (albeit, a large one). All prophages present in EHEC O157 alter its potential for recombination and therefore may alter its phenotype and virulence potential in the process (**Chapter 4**). Furthermore, the concepts of an IS mediated phage entrapment and Stx cassettes were discussed (**Chapters 2** and **3**). When combining all these concepts and observations together, one can propose an “ecosystem” of prophages, where different prophages impact on the phentoype of the host bacterium in different ways, or at least impact on phenotype by different combinations of mechanisms.

Based on this concept, one can loosely classify the role of a prophage within three classes:

1. “direct transcriptomic” (when the prophage expresses a protein that affects the bacterium or host)
2. “indirect transcriptomic” (when the prophage mediates LCRs, reshuffling the prophage genetic content, resulting in a different transcriptomic pattern)
3. “genomic disruption” (when the prophage variation, mainly due to IS elements, disrupts gene expression or prophage induction and excision)

This classification is not only supported by this work but also by the literature. The “direct transcriptomic” aspect is the one most understood, having been studied in depth as part of horizontal gene transfer, where a prophage is only worth the genes it contains that can be expressed (such as *stx*) (Asadulghani et al. 2009; Ogura et al. 2015). LCRs presented in the concept of “indirect transcriptomics” have been seen in other organisms. *Yersinia pestis* being a prime example, exhibiting a large amount of LCRs as shown by Darling *et al*, which had similar observations, such as the importance of the origin of replication in this process (Darling, Miklós, and Ragan 2008). Furthermore, other works hypothesised that IS elements are involved in LCRs in *E. coli* (Raeside et al. 2014; Lee et al. 2016). However, the Raeside *et al* (Raeside et al. 2014) methodology used optical mapping and short-read sequencing; therefore, it could be argued that it may not be solely the IS elements at play but the surrounding areas of homology that were described in this work. Our work supports this argument with the observed LCRs correlating with larger areas of homology. The Lee *et al* (Lee et al. 2016) body of work further confirms this, as while it focused on the rates of IS elements insertions and LCRs, they do also find deletions that were not IS mediated. Therefore, it is clear that IS elements are not the sole cause of LCRs. The concept of “genomic disruption” has been less observed due to IS elements being hard to resolve prior to long-read sequencing. However, work from Asaldughani *et*

al (Asadulghani et al. 2009) has shown that a proportion of the EHEC O157 prophage pool cannot be induced due to non-functional excisionases, sometimes even potentially requiring the mechanism of two prophages to achieve it.

Therefore, the final model proposed through this work involves: typical horizontal gene transfer through plasmids and bacteriophages mediating isolate and phenotype variation, large areas of homologies and IS elements mediating LCRs allowing for a single bacterium to possess a potential for plasticity and generate different genomic “configuration”, and IS mediated gene disruptions affecting specific gene expression and potential prophage induction, with the potential of returning the gene integrity upon the excision of the IS element

All the above allows for different phenotypes to be exhibited rapidly without the need to rely on external genetic content. This level of plasticity would allow for niche expansions and rapid response to varying conditions. This could also explain why the PFGE profiles observed in clinical samples exhibited a level of stability in profile C (**Chapter 4**), which is most likely being a very successful genomic “configuration” for human colonization.

5.2 Limitations of Study and Future Works

As previously mentioned, the key limitation of this work is the availability of long-read EHEC O157 whole genome sequences, which can also be attributable to some of the other limitations. This work was started in September 2014 when there were only four widely used EHEC O157 reference whole genome sequences (Sakai, TW14359, EC4115, and EDL933) (respective accession numbers: BA000007, NC_013008, NC_011353, and NC_002655). Therefore the addition of 69 whole genome sequences, some of which have known relationships, is a great new source of data. However to exhaustively investigate the hypothesis and claims made throughout this study, a much larger sequence pool, with a wider variety of relationship levels, needs to be obtained. The limited number of sequences has led this work to be an observational study, laying the foundation for a large amount of potential future work. This was further compounded by the lack of short-read sequencing data for the majority of these 69 isolates, complicating the extrapolation of data. The lack of traditional typing results also caused issues as none of the isolates originating from the USA had a PT, and only a small subset of all isolates had a determined PFGE profile. Finally, metadata on the isolates was also extremely sparse.

While having typing data from prior techniques may seem obsolete, it allows for the question being investigated to be focused. For example, one could determine the effect of LCRs on PFGE type and whether outbreaks that were called using this method might have been inaccurately typed. This seems highly likely given that PFGE profiles are determined through generation of band sizes resulting from the genome being digested at specific restriction sites. Therefore, an LCR has a high likelihood of changing that profile, which is what was observed (**Chapter 5**). On the other hand, comparing other typing methods such as phage typing to whole genome arrangement might offer novel insight into biological mechanisms of the pathogen, as well as potentially establish null hypotheses for novel response techniques such as phage therapy. Phage therapy consists of treating a bacterial infection by

targeting it using bacteriophages to which the bacterium is known to be susceptible. Therefore understanding how phage susceptibility and resistance work through studying phage typing, and the potential effects of LCRs on this mechanism, could lead to useful results.

Furthermore it may seem counter-intuitive to require short-read sequencing data for these isolates. However having this data available allows for findings to be extrapolated and “translated” in a way that allows their use in the PH field, which currently relies on short-read technology platforms. A recent paper from Greig *et al* (Greig et al. 2019) observes the differences between Illumina short-read and Oxford Nanopore long-read SNP calling. They found minor variations in results that indicate and conclude that both methodologies probably have a degree of false calls, but that the conclusions and relationships observed remain constant (Greig et al. 2019). Therefore, in works such as those presented here, short-read sequencing is required to validate and verify results obtained by long-read sequencing until these differences are fully understood.

Finally, the lack of metadata is one of the largest data limitations. Having metadata allows us to conduct association studies, linking genomic observations to specific phenotypes. For example, with the right metadata one can link specific genes or mutations to antimicrobial resistance, virulence, clinical outcome, host age, and symptom severity. If metadata was available, LCRs could potentially be associated with such phenotypes. This would, be quite a costly experiment to run. Ideally one would need to use long-read sequencing on whole outbreak events and obtain corresponding clinical data which in turn requires permission by ethics and data governance committees.

Another set of limitations is technological, as mentioned in **Section 4.4.3**, where to truly investigate LCRs one would need to investigate single cell events, or genomic frequencies within a population. Furthermore, this work is also limited by the rather minimal information and reference data available on prophages. Due to prior limitation in sequencing prophages using short-read

sequencing platforms, the amount of data on the subject is still relatively low. When looking at the tools that were used for prophage calling in this study: PHAST and PHASTER (Zhou et al. 2011; Arndt et al. 2016), both offer a large database of prophage associate genes, and PHASTER was released while this study was ongoing. However, the results, while similar, are not identical, showing that the understanding of prophages and their boundaries is still evolving with novel, more efficient methods still being developed (**Appendix II-A**). Therefore these types of studies will only increase in impact and confidence as more prophages, and EHEC O157, isolates are fully sequenced and assembled.

There were also limitations due to available data regarding IS elements. An observation made in **Section 4.4.1** discusses how certain IS elements may not have been detected due to being reliant on curation of IS databases which are still expanding. Furthermore, the same observation leads to the possibility of IS elements not being the sole mechanism behind LCRs. However, until relatively recently “long-read” sequencing was not widely available to fully assemble genome sections with large numbers of IS elements. This should allow for more genomes to be generated with unambiguous assemblies. With this level of resolution it becomes possible to attempt and determine exact boundaries of LCRs and whether IS elements are within those boundaries or simply nearby (as done in **Section 4.4.1**). However, this should be done in a more consistent manner, whilst looking at a larger number of LCRs of different types rather than the limited examples done in this work.

The reason only limited examples were looked at is in fact the key limitation of this work, time. With 69 sequenced isolates, the amount of data generated was considerable and was amplified by the fact that in most cases, pair-wise analysis needed to be conducted rather than population wide studies due to the diversity present in the population. This limitation is partly the reason why most of the study was observational. To conclusively prove any of the work presented here, one would need to focus greatly on a specific subset of the

samples, or most likely develop a novel method to analyse the whole data, as specific sampling was observed to be a limitation at times in this work. Machine learning exhibits potential for this but has the drawback of being a “black box” system, where the reason behind results can be difficult to determine or understand fully. The field of Bioinformatics has for a long time mostly developed techniques adapted for short-read sequencing, such as kmer hashing which has become widespread in the investigation of a myriad of questions. However, with the increasing popularity of MinIon, novel methods are being developed to better utilise the potential of long-read sequencing. This, in conjunction with the ever-growing field of AI and machine learning should allow for better analysis of larger datasets. Consequently, new methods will develop, allowing for the further work introduced in this section to be conducted more efficiently.

5.3 The Impact of Long-Read Sequencing on Public Health

The new role WGS is finding in PH was discussed in **Chapter 1**. However, this was prior to the data generated from long-read sequencing being introduced. This section will explore how long-read sequencing and the data that can be generated from it can supplement PH, if at all.

As previously described, the main sequencing technology platform used in PH is the Illumina “short-read” sequencing platform. However, the MinIon platform has been used due to its portability and simple DNA preparation process, for the study of viral epidemics such as the Ebola and Zika outbreaks that were seen in 2015 and 2017 respectively, in Sierra Leone and Brazil respectively (Quick et al. 2016; 2017). In both these cases the challenges faced were different than the ones presented in **Chapter 1**. WGS for viral outbreaks can be used to identify the organism, determine its rate of evolution as well as other features such as further insight in the viral response to vaccines or treatments. All this data allows for outbreak surveillance and can greatly supplement the field. However, these are the advantages of WGS in general, not necessarily “long-read” sequencing. The MinIon offers key advantages that cannot be obtained by any other platform. It is portable and requires minimal lab equipment. Considering that these two outbreaks were in relatively less developed countries, these advantages were key to supply WGS data to the PH institutions and improve epidemiological and clinical responses (Quick et al. 2017; 2016).

While the previous example made use of the platform, it didn’t fully utilize its potential as a long-read sequencing platform. Recently “long-read” sequencing platforms have been used in specific cases to observe human genetic conditions (Ardui et al. 2018). Such works utilised long-read sequencing to solve assembly breakage due to repeated regions that could not be resolved through “short-read” sequencing, similarly to how EHEC O157 prophages and genomes were difficult to fully assemble prior to “long-read” sequencing technologies. On the microbial side, a lot of novel work is

focusing on AMR and plasmids (González-Escalona et al. 2019). Mobile genomic elements (such as plasmids) are being further studied due to their potentially high impact on phenotype, and the lesser difficulty in their study using “long-read” sequencing. Focusing, on EHEC O157, this is even more relevant as the mobile genetic elements also include prophages, which make up a considerable proportion of the EHEC O157 genomic material. As shown in this work, understanding the prophage content can supplement the phylogenomic data when metadata and “short-read” sequencing may be ambiguous (**Section 3.4.1**). However, more importantly, this type of work could truly be beneficial to PH if LCRs are found to have a large impact on phenotype *in vivo*. Not only would this data then allow to determine the direct risk an isolate poses, but also its potential virulence for different configuration thus allowing for potential prioritisation of outbreak management.

While this is based on the hypothesis presented in this study, it is an even more unlikely hypothesis if “long-read” sequencing does not get fully adopted by the PH field. The main barriers to PH are costs and time. While these may appear to be straightforward metrics, a lot of thought is required when considering them. Cost is not necessarily selecting the cheapest option, but the one that saves the PH field the most amount of money. Therefore, an expensive option that could help prevent 75% of cases (with each case having its own cost) would be better than a cheaper option which only help prevent 25% of cases, given that the difference in cost between the two methods is less than the savings made by preventing that supplementary 50% of cases. Time is estimated in the same manner, a test with a quick turnaround is preferred, but is only as good as the utility of the information it generates. Currently, the Illumina platforms offer the best value for money. However, as previously discussed, MinIon sequencing is starting to make a case for itself with its decreasing cost, simpler DNA preparation protocol, long reads, and improving error rate. On an optimized workflow, MinIon sequencing was found to be as cheap as £60-£80 per sample by our PHE collaborators. However, this was to supplement Illumina sequencing, which costs £60 per sample (Scottish Reference Laboratories, Personal

communication) and therefore, is probably cheaper than it would be if used by itself. However, considering how much higher the cost of “long-read” sequencing was less than a year ago, this is incredible improvement. Furthermore, recent developments by the PHE team found the use of MinIon sequencing to offer other advantages (Greig et al. 2019). Due to the ability of the MinIon to allow for live base calling, this allows for certain basic typing results (such as serotype, MLST and toxin typing) to be available while the sequencing run is still ongoing and more rapidly than when performing Illumina short-read sequencing (Greig et al. 2019). This would eventually permit long-read sequencing to be a valid replacement to rapid wet lab typing methodologies.

As previously stated, cost is relative to the utility of the data generated. Therefore, as more academic work investigates the methylation profiles that can be generated using the PacBio platform and their implication on phenotype, the more likely it is to become an interesting test for PH if it is of clinical relevance. Finally, the future might be a hybrid solution given that Illumina just recently acquired PacBio. Therefore, while it is unclear what the future holds, it seems unlikely that “long-read” sequencing and the further insight it generates will not be part of it., This work will hopefully serve as a stepping stone to guide future works in this area.

6 Bibliography

Afset, J. E., E. Anderssen, G. Bruant, J. Harel, L. Wieler, and K. Bergh. 2008. "Phylogenetic Backgrounds and Virulence Profiles of Atypical Enteropathogenic Escherichia Coli Strains from a Case-Control Study Using Multilocus Sequence Typing and DNA Microarray Analysis." *Journal of Clinical Microbiology* 46 (7): 2280–90. https://doi.org/10.1128/JCM.01752-07.

Ahmed, R., C. Bopp, A. Borczyk, and S. Kasatiya. 1987. "Phage-Typing Scheme for Escherichia Coli O157:H7." *The Journal of Infectious Diseases* 155 (4): 806–9. https://doi.org/10.1093/infdis/155.4.806.

Aird, D., M. G. Ross, WS. Chen, M. Danielsson, T. Fennell, C. Russ, D. Jaffe, C. Nusbaum, and A. Gnirke. 2011. "Analyzing and Minimizing PCR Amplification Bias in Illumina Sequencing Libraries." *Genome Biology* 12 (2): R18–R18. https://doi.org/10.1186/gb-2011-12-2-r18.

Akashi, S., K. Joh, A. Tsuji, H. Ito, H. Hoshi, T. Hayakawa, J. Ihara, et al. 1994. "A Severe Outbreak of Haemorrhagic Colitis and Haemolytic Uraemic Syndrome Associated with Escherichia Coli 0157:H7 in Japan." *Eur J Pediatr* 153. https://doi.org/10.1007/BF02190685.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Ardui, S., A. Ameur, J. R. Vermeesch, and M. S. Hestand. 2018. "Single Molecule Real-Time (SMRT) Sequencing Comes of Age: Applications and Utilities for Medical Diagnostics." *Nucleic Acids Research* 46 (5): 2159–68. https://doi.org/10.1093/nar/gky066.

Armstrong, . L., J. Hollingsworth, and J. G. Morris. 1996. "Emerging Foodborne Pathogens: Escherichia Coli O157:H7 as a Model of Entry of a New Pathogen into the Food Supply of the Developed World." *Epidemiologic Reviews* 18 (1): 29–51. https://doi.org/10.1093/oxfordjournals.epirev.a017914.

Arndt, D., J. R. Grant, A. Marcu, T. Sajed, A. Pon, Y. Liang, and D. S. Wishart.

2016. “PHASTER: A Better, Faster Version of the PHAST Phage Search Tool.” *Nucleic Acids Research* 44 (W1): W16–21. https://doi.org/10.1093/nar/gkw387.

Asadulghani, M., Y. Ogura, T. Ooka, T. Itoh, A. Sawaguchi, A. Iguchi, K. Nakayama, and T. Hayashi. 2009. “The Defective Prophage Pool of Escherichia Coli O157: Prophage-Prophage Interactions Potentiate Horizontal Transfer of Virulence Determinants.” *PLoS Pathog* 5 (5): e1000408. https://doi.org/10.1371/journal.ppat.1000408.

Ashton, P. M., N. Perry, R. Ellis, L. Petrovska, J. Wain, K. A. Grant, C. Jenkins, and T. J. Dallman. 2015. “Insight into Shiga Toxin Genes Encoded by Escherichia Coli O157 from Whole Genome Sequencing.” *PeerJ* 3: e739. https://doi.org/10.7717/peerj.739.

Balasubramanian, S., M.S. Osburne, H. BrinJones, A. K. Tai, and J. M. Leong. 2019. “Prophage Induction, but Not Production of Phage Particles, Is Required for Lethal Disease in a Microbiome-Replete Murine Model of Enterohemorrhagic E. Coli Infection.” *PLOS Pathogens* 15 (1): e1007494. https://doi.org/10.1371/journal.ppat.1007494.

Barquist, L., M. Mayho, C. Cummins, A. K. Cain, C. J. Boinett, A. J. Page, G. C. Langridge, M. A. Quail, J. A. Keane, and J. Parkhill. 2016. “The TraDIS Toolkit: Sequencing and Analysis for Dense Transposon Mutant Libraries.” *Bioinformatics (Oxford, England)* 32 (7): 1109–11. https://doi.org/10.1093/bioinformatics/btw022.

Belén, Ana, Ibarz Pavón, and Martin C J Maiden. 2009. “Molecular Sequence Typing.” *Methods in Molecular Biology* 551 (8): 129–40. https://doi.org/10.1007/978-1-60327-999-4.

Betz, J., M. Bielaszewska, A. Thies, HU. Humpf, K. Dreisewerd, H. Karch, K. S. Kim, A. W. Friedrich, and J. Müthing. 2011. “Shiga Toxin Glycosphingolipid Receptors in Microvascular and Macrovascular Endothelial Cells: Differential Association with Membrane Lipid Raft Microdomains.” *Journal of Lipid Research* 52 (4): 618–34. https://doi.org/10.1194/jlr.M010819.

Blanco, M., J. E. Blanco, J. Blanco, E. A. Gonzalez, A. Mora, C. Prado, L.

Fernandez, M. Rio, J. Ramos, and M. P. Alonso. 1996. “Prevalence and Characteristics of Escherichia Coil Serotype O157:H7 and Other Verotoxin-Producing E. Coli in Healthy Cattle.” *Epidemiology and Infection* 117 (2): 251–57. https://doi.org/DOI: 10.1017/S0950268800001424.

Byrne, L., R. Elson, T. J. Dallman, N. Perry, P. Ashton, J. Wain, G. K. Adak, K. A. Grant, and C. Jenkins. 2014. “Evaluating the Use of Multilocus Variable Number Tandem Repeat Analysis of Shiga Toxin-Producing Escherichia Coli O157 as a Routine Public Health Tool in England.” *PLoS ONE* 9 (1). https://doi.org/10.1371/journal.pone.0085901.

Byrne, L., C. Jenkins, N. Launders, R. Elson, and G. K. Adak. 2015. “The Epidemiology, Microbiology and Clinical Impact of Shiga Toxin-Producing Escherichia Coli in England, 2009–2012.” *Epidemiology and Infection* 143 (16): 3475–87. https://doi.org/DOI: 10.1017/S0950268815000746.

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

Chase-Topping, Margo, David Gally, Chris Low, Louise Matthews, and Mark Woolhouse. 2008. “Super-Shedding and the Link between Human Infection and Livestock Carriage of Escherichia Coli O157.” *Nature Reviews. Microbiology* 6 (12): 904–12. https://doi.org/10.1038/nrmicro2029.

Chin, C. S., D. Alexander, P. Marks, A. Klammer, J. Drake, C. Heiner, A. Clum, et al. 2013. “Nonhybrid, Finished Microbial Genome Assemblies from Long-Read SMRT Sequencing Data.” *Nature Methods* 10 (6). https://doi.org/10.1038/nmeth.2474.

Cowley, L. A., T. J. Dallman, S. Fitzgerald, N. Irvine, P. J. Rooney, S. P. McAteer, M. Day, et al. 2016. “Short-Term Evolution of Shiga Toxin-Producing Escherichia Coli O157:H7 between Two Food-Borne Outbreaks.” *Microbial Genomics* 2 (9): e000084. https://doi.org/10.1099/mgen.0.000084.

Cunningham, S. A., N. Chia, P. R. Jeraldo, D. J. Quest, J. A. Johnson, D. J. Boxrud, A. J. Taylor, et al. 2017. “Comparison of Whole-Genome Sequencing Methods

for Analysis of Three Methicillin-Resistant Staphylococcus Aureus Outbreaks." *Journal of Clinical Microbiology* 55 (6): 1946–53. https://doi.org/10.1128/JCM.00029-17.

Dallman, T. J., P. M. Ashton, L. Byrne, N. Perry, L. Petrovska, R. Ellis, L. Allison, et al. 2015. "Applying Phylogenomics to Understand the Emergence of Shiga-Toxin-Producing Escherichia Coli O157:H7 Strains Causing Severe Human Disease in the UK." *Microbial Genomics*. https://doi.org/10.1099/mgen.0.000029.

Dallman, T. J., P. Ashton, U. Schafer, A. Jironkin, A. Painset, S. Shaaban, H. Hartman, et al. 2018. "SnapperDB: A Database Solution for Routine Sequencing Analysis of Bacterial Isolates." *Bioinformatics* 34 (17): 3028–29. https://doi.org/10.1093/bioinformatics/bty212.

Darling, A. E., B. Mau, and N. T. Perna. 2010. "Progressivemauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement." *PLoS ONE* 5 (6). https://doi.org/10.1371/journal.pone.0011147.

Darling, A. E., I. Miklós, and M. A. Ragan. 2008. "Dynamics of Genome Rearrangement in Bacterial Populations." *PLOS Genetics* 4 (7): e1000128. https://doi.org/10.1371/journal.pgen.1000128.

Delcher, A. L., K. A. Bratke, E. C. Powers, and S. L. Salzberg. 2007. "Identifying Bacterial Genes and Endosymbiont DNA with Glimmer." *Bioinformatics* 23 (6): 673–79. https://doi.org/10.1093/bioinformatics/btm009.

Eppinger, M., M. K. Mammel, J. E. Leclerc, J. Ravel, and T. A. Cebula. 2011. "Genomic Anatomy of Escherichia Coli O157:H7 Outbreaks." *Proceedings of the National Academy of Sciences* 108 (50): 20142–47. https://doi.org/10.1073/pnas.1107176108.

Feng, P., and S. Monday. 2000. "Multiplex PCR for Detection of Trait and Virulence Factors in Enterohemorrhagic Escherichia Coli Serotypes." *Molecular and Cellular Probes* 14 (6): 333–37. https://doi.org/10.1006/mcpr.2000.0323.

Fijalkowska, I. J., R. M. Schaaper, and P. Jonczyk. 2012. "DNA Replication Fidelity in Escherichia Coli: A Multi-DNA Polymerase Affair." *FEMS Microbiology*

Reviews 36 (6): 1105–21. https://doi.org/10.1111/j.1574-6976.2012.00338.x.

Fitzgerald, S. F., A. E. Beckett, J. Palarea-Albaladejo, S. McAteer, S. Shaaban, J. Morgan, N. I. Ahmad, et al. 2019. "Shiga Toxin Sub-Type 2a Increases the Efficiency of Escherichia Coli O157 Transmission between Animals and Restricts Epithelial Regeneration in Bovine Enteroids." *PLoS Pathogens* 15 (10). https://doi.org/10.1371/journal.ppat.1008003.

Fratamico, P. M., S. K. Sackitey, M. Wiedmann, and Y. D. Ming. 1995. "Detection of Escherichia Coli O157:H7 by Multiplex PCR." *Journal of Clinical Microbiology* 33 (8): 2188–91.

Fuller, C. A., C. A. Pellino, M. J. Flagler, J. E. Strasser, and A. A. Weiss. 2011. "Shiga Toxin Subtypes Display Dramatic Differences in Potency." Edited by S R Blanke. *Infection and Immunity* 79 (3): 1329 LP – 1337. https://doi.org/10.1128/IAI.01182-10.

Geneious. 2019. "Geneious." 2019. https://www.geneious.com/. Accessed 01/09/2019. Web Archive: https://web.archive.org/web/20190906041414/https://www.geneious.com/.

González-Escalona, N., M. A. Allard, E. W. Brown, S. Sharma, and M. Hoffmann. 2019. "Nanopore Sequencing for Fast Determination of Plasmids, Phages, Virulence Markers, and Antimicrobial Resistance Genes in Shiga Toxin-Producing Escherichia Coli." *Plos One* 14 (7): e0220494. https://doi.org/10.1371/journal.pone.0220494.

Greig, D. R., C. Jenkins, S. Gharbia, and T. J. Dallman. 2019. "Comparison of Single-Nucleotide Variants Identified by Illumina and Oxford Nanopore Technologies in the Context of a Potential Outbreak of Shiga Toxin–Producing Escherichia Coli." *GigaScience* 8 (8). https://doi.org/10.1093/gigascience/giz104.

Hänninen, M. L., M. Hakkinen, and H. Rautelin. 1999. "Stability of Related Human and Chicken Campylobacter Jejuni Genotypes after Passage through Chick Intestine Studied by Pulsed-Field Gel Electrophoresis." *Applied and Environmental Microbiology* 65 (5): 2272–75.

https://www.ncbi.nlm.nih.gov/pubmed/10224037.

Ho, N. K., A. C. Henry, K. Johnson-Henry, and P.M. Sherman. 2013. "Pathogenicity, Host Responses and Implications for Management of Enterohemorrhagic Escherichia Coli O157:H7 Infection." *Canadian Journal of Gastroenterology = Journal Canadien de Gastroenterologie* 27 (5): 281–85. https://doi.org/10.1155/2013/138673.

Holmes, A., T. J. Dallman, S. Shabaan, M. Hanson, and L. Allison. 2018. "Validation of Whole-Genome Sequencing for Identification and Characterization of Shiga Toxin-Producing Escherichia Coli To Produce Standardized Data To Enable Data Sharing." Edited by Alexander Mellmann. *Journal of Clinical Microbiology* 56 (3): e01388-17. https://doi.org/10.1128/JCM.01388-17.

Hyatt, D., GL. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, and L. J. Hauser. 2010. "Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification." *BMC Bioinformatics* 11 (March): 119. https://doi.org/10.1186/1471-2105-11-119.

Illumina, Inc. 2019. "Illumina Platforms." 2019. https://www.illumina.com/systems/sequencing-platforms.html. Accessed 31/08/2019. Web Archive: https://web.archive.org/web/20191006190049/https://www.illumina.com/systems/sequencing-platforms.html.

Imamovic, L., R. Tozzoli, V. Michelacci, F. Minelli, M. L. Marziano, A. Caprioli, and S. Morabito. 2010. "OI-57, a Genomic Island of Escherichia Coli O157, Is Present in Other Seropathotypes of Shiga Toxin-Producing E. Coli Associated with Severe Human Disease." *Infection and Immunity* 78 (11): 4697 LP – 4704. https://doi.org/10.1128/IAI.00512-10.

Inkscape. 2019. "InkScape." 2019. https://inkscape.org/. Accessed 31/08/2019. Web Archive: https://web.archive.org/web/20190915014942/https://inkscape.org/.

Izumiya, H., Y. Pei, J. Terajima, M. Ohnishi, T. Hayashi, S. Iyoda, and H. Watanabe. 2010. "New System for Multilocus Variable-Number Tandem-Repeat Analysis

of the Enterohemorrhagic Escherichia Coli Strains Belonging to Three Major Serogroups: O157, O26, and O111." *Microbiology and Immunology* 54 (10): 569–77. https://doi.org/10.1111/j.1348-0421.2010.00252.x.

Izumiya, H., J. Terajima, A. Wada, Y. Inagaki, K. I. Itoh, K. Tamura, and H. Watanabe. 1997. "Molecular Typing of Enterohemorrhagic Escherichia Coli O157:H7 Isolates in Japan by Using Pulsed-Field Gel Electrophoresis." *Journal of Clinical Microbiology* 35 (7): 1675–80.

Janowicz, A., F. De Massis, M. Ancora, C. Cammà, C. Patavino, A. Battisti, K. Prior, et al. 2018. "Core Genome Multilocus Sequence Typing and Single Nucleotide Polymorphism Analysis in the Epidemiology of Brucella Melitensis Infections." Edited by Daniel J Diekema. *Journal of Clinical Microbiology* 56 (9): e00517-18. https://doi.org/10.1128/JCM.00517-18.

Kaguni, J.M. 2011. "Replication Initiation at the Escherichia Coli Chromosomal Origin." *Current Opinion in Chemical Biology* 15 (5): 606–13. https://doi.org/10.1016/j.cbpa.2011.07.016.

Kaper, J., J. Nataro, and H. Mobley. 2004. "Pathogenic Escherichia Coli." *Nat Rev Microbiol* 2.

Karch, H., E. Denamur, U. Dobrindt, B. B. Finlay, R. Hengge, L. Johannes, E. Z. Ron, T. Tønjum, P. J. Sansonetti, and M. Vicente. 2012. "The Enemy within Us: Lessons from the 2011 European Escherichia Coli O104:H4 Outbreak." *EMBO Molecular Medicine* 4 (9): 841–48. https://doi.org/10.1002/emmm.201201662.

Katani, R., R. Cote, J. A. R. Garay, L. Li, T. M. Arthur, C. DebRoy, M. M. Mwangi, and V. Kapur. 2016. "Complete Genome Sequence of SS52, a Strain of Escherichia Coli O157: H7 Recovered from Supershedder Cattle." *Genome Announcements* 3 (2): 1999–2000. https://doi.org/10.1128/genomeA.01569-14.

Keim, P., L. B. Price, A. M. Klevytska, K. L. Smith, J. M. Schupp, R. Okinaka, P. J. Jackson, and M. E. Hugh-Jones. 2000. "Multiple-Locus Variable-Number Tandem Repeat Analysis Reveals Genetic Relationships within Bacillus Anthracis." *Journal of Bacteriology* 182 (10): 2928–36.

https://doi.org/10.1128/JB.182.10.2928-2936.2000.

Khakhria, R., D. Duck, and H. Lior. 1990. “Extended Phage-Typing Scheme for Escherichia Coli 0157:H7.” *Epidemiology and Infection* 105 (3): 511–20. https://doi.org/10.1017/S0950268800048135.

Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy. 2017. “Canu: Scalable and Accurate Long-Read Assembly via Adaptive k-Mer Weighting and Repeat Separation.” *Genome Research* 27 (5): 722–36. https://doi.org/10.1101/gr.215087.116.

Krzywinski, M.I., J. E. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. 2009. “Circos: An Information Aesthetic for Comparative Genomics.” *Genome Research*, June. https://doi.org/10.1101/gr.092759.109.

Kulasekara, B. R., M. Jacobs, Y. Zhou, Z. Wu, E. Sims, C. Saenphimmachak, L. Rohmer, et al. 2009. “Analysis of the Genome of the Escherichia Coli O157:H7 2006 Spinach-Associated Outbreak Isolate Indicates Candidate Genes That May Enhance Virulence.” *Infection and Immunity* 77 (9): 3713–21. https://doi.org/10.1128/IAI.00198-09.

Laing, C. R., Y. Zhang, M.W. Gilmour, V. Allen, R. Johnson, J. E. Thomas, and V. P. J. Gannon. 2012. “A Comparison of Shiga-Toxin 2 Bacteriophage from Classical Enterohemorrhagic Escherichia Coli Serotypes and the German E. Coli O104:H4 Outbreak Strain.” *PLoS ONE* 7 (5). https://doi.org/10.1371/journal.pone.0037362.

Lee, H., T.G. Doak, E. Popodi, P. L. Foster, and H. Tang. 2016. “Insertion Sequence-Caused Large-Scale Rearrangements in the Genome of Escherichia Coli.” *Nucleic Acids Research* 44 (15): 7109–19. https://doi.org/10.1093/nar/gkw647.

Li, H., and R. Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60. https://doi.org/10.1093/bioinformatics/btp324.

Lim, J.Y., J. Yoon, and C. J. Hovde. 2010. “A Brief Overview of Escherichia Coli O157:H7 and Its Plasmid O157.” *Journal of Microbiology and Biotechnology*

20 (1): 5–14. https://www.ncbi.nlm.nih.gov/pubmed/20134227.

Luo, H., CT. Zhang, and F. Gao. 2014. “Ori-Finder 2, an Integrated Tool to Predict Replication Origins in the Archaeal Genomes.” *Frontiers in Microbiology* 5: 482. https://doi.org/10.3389/fmicb.2014.00482.

Maharjan, R. P., and T. Ferenci. 2018. “The Impact of Growth Rate and Environmental Factors on Mutation Rates and Spectra in Escherichia Coli.” *Environmental Microbiology Reports* 10 (6): 626–33. https://doi.org/10.1111/1758-2229.12661.

Maiden, M. C. J. 2006. “Multilocus Sequence Typing of Bacteria.” *Annual Review of Microbiology* 60 (1): 561–88. https://doi.org/10.1146/annurev.micro.59.030804.121325.

Maiden, M. C. J., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, et al. 1998. “Multilocus Sequence Typing: A Portable Approach to the Identification of Clones within Populations of Pathogenic Microorganisms (Molecular TypingNeisseria Meningitidishousekeeping GenesWorld-Wide Webhyper-Virulent Clones).” *Proc. Natl. Acad. Sci. USA* 95 (March): 3140–45. www.pnas.org.

Matthews, T. D., and S. Maloy. 2010. “Fitness Effects of Replichore Imbalance in Salmonella Enterica.” *Journal of Bacteriology* 192 (22): 6086 LP – 6088. https://doi.org/10.1128/JB.00649-10.

Mcdaniel, T. K., K. G. Jarvis, M. S. Donnenberg, and J. B. Kaper. 1995. “A Genetic Locus of Enterocyte Effacement Conserved among Diverse Enterobacterial Pathogens.” *Proceedings of the National Academy of Sciences of the United States of America* 92 (5): 1664–68. https://doi.org/10.1073/pnas.92.5.1664.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” *Genome Res* 20 (9): 1297–1303. https://doi.org/10.1101/gr.107524.110.

Melton-Celsa, A. R. 2014. “Shiga Toxin (Stx) Classification, Structure, and Function.” *Microbiology Spectrum* 2 (4): 10.1128/microbiolspec.EHEC-0024-

2013–2013. https://doi.org/10.1128/microbiolspec.EHEC-0024-2013.

Mikheyev, A. S., and M. M. Tin. 2014. “A First Look at the Oxford Nanopore MinION Sequencer.” *Mol Ecol Resour* 14 (6): 1097–1102. https://doi.org/10.1111/1755-0998.12324.

Nadon, C. A., E. Trees, L. K. Ng, E. Møller Nielsen, A. Reimer, N. Maxwell, K. A. Kubota, P. Gerner-Smidt, and MLVA Harmonization Working Group. 2013. “Development and Application of MLVA Methods as a Tool for Inter-Laboratory Surveillance.” *Euro Surveillance : Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* 18 (35): 20565. https://doi.org/10.2807/1560-7917.es2013.18.35.20565.

Needleman, S. B., and C. D. Wunsch. 1970. “A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins.” *Journal of Molecular Biology* 48 (3): 443–53. https://doi.org/https://doi.org/10.1016/0022-2836(70)90057-4.

Neylon, C., A. V. Kralicek, T. M. Hill, and N. E. Dixon. 2005. “Replication Termination in Escherichia Coli: Structure and Antihelicase Activity of the Tus-Ter Complex.” *Microbiology and Molecular Biology Reviews : MMBR* 69 (3): 501–26. https://doi.org/10.1128/MMBR.69.3.501-526.2005.

O’Brien, A. D., J. W. Newland, S. F. Miller, R. K. Holmes, H. W. Smith, and S. B. Formal. 1984. “Shiga-like Toxin-Converting Phages from Escherichia Coli Strains That Cause Hemorrhagic Colitis or Infantile Diarrhea.” *Science* 226 (4675): 694 LP – 696. https://doi.org/10.1126/science.6387911.

Ofir, G., and R. Sorek. 2018. “Contemporary Phage Biology: From Classic Models to New Insights.” *Cell* 172 (6): 1260–70. https://doi.org/10.1016/j.cell.2017.10.045.

Ogura, Y., S. I. Mondal, M. R. Islam, T. Mako, K. Arisawa, K. Katsura, T. Ooka, et al. 2015. “The Shiga Toxin 2 Production Level in Enterohemorrhagic Escherichia Coli O157:H7 Is Correlated with the Subtypes of Toxin-Encoding Phage.” *Sci Rep* 5: 16663. https://doi.org/10.1038/srep16663.

Ohnishi, M, K Kurokawa, and T Hayashi. 2001. “Diversification of Escherichia Coli

Genomes: Are Bacteriophages the Major Contributors?" *Trends in Microbiology* 9 (10).

Ooka, T., Y. Ogura, M. Asadulghani, M. Ohnishi, K. Nakayama, J. Terajima, H. Watanabe, and T. Hayashi. 2009. "Inference of the Impact of Insertion Sequence (IS) Elements on Bacterial Genome Diversification through Analysis of Small-Size Structural Polymorphisms in Escherichia Coli O157 Genomes." *Genome Research* 19. https://doi.org/10.1101/gr.089615.108.

Oxford Nanopore. 2017. "Nanopore 1Mbp Read." 2017. https://nanoporetech.com/about-us/news/world-first-continuous-dna-sequence-more-million-bases-achieved-nanopore-sequencing. Accessed 31/08/2019. Web Archive: NA.

———. 2019. "Oxford Nanopore." 2019. https://nanoporetech.com/. Accessed 31/08/2019. Web Archive: https://web.archive.org/web/20190902104823/https://nanoporetech.com/.

Pacheco, A. R., and V. Sperandio. 2012. "Shiga Toxin in Enterohemorrhagic E.Coli: Regulation and Novel Anti-Virulence Strategies." *Frontiers in Cellular and Infection Microbiology* 2 (June): 81. https://doi.org/10.3389/fcimb.2012.00081.

Pacific Biosciences. 2019. "PacBio." 2019. https://www.pacb.com/. Accessed: 31/08/2019. Web Archive: https://web.archive.org/web/20190905232920/https://www.pacb.com/.

Pearce, M. E., NF. Alikhan, T. J. Dallman, Z. Zhou, K. Grant, and M. C. J. Maiden. 2018. "Comparative Analysis of Core Genome MLST and SNP Typing within a European Salmonella Serovar Enteritidis Outbreak." *International Journal of Food Microbiology* 274 (June): 1–11. https://doi.org/10.1016/j.ijfoodmicro.2018.02.023.

Pei, Y., J. Terajima, Y. Saito, R. Suzuki, N. Takai, H. Izumiya, T. Morita-Ishihara, et al. 2008. "Molecular Characterization of Enterohemorrhagic Escherichia Coli O157:H7 Isolates Dispersed across Japan by Pulsed-Field Gel Electrophoresis and Multiple-Locus Variable-Number Tandem Repeat Analysis." *Japanese Journal of Infectious Diseases* 61 (1): 58–64.

Perna, N., G. Plunkett III, V. Burland, B. Mau, J. Glasner, D. Rose, G. Mayhew, et al. 2001. “Genome Sequence of Enterohaemorrhagic Escherichia Coli O157:H7.” *Nature* 409. https://doi.org/10.1038/35054089.

Persson, S., K. E. P. Olsen, S. Ethelberg, and F. Scheutz. 2007. “Subtyping Method for Escherichia Coli Shiga Toxin (Verocytotoxin) 2 Variants and Correlations to Clinical Manifestations.” *Journal of Clinical Microbiology* 45 (6): 2020–24. https://doi.org/10.1128/JCM.02591-06.

PHE Bioinformatics Unit. 2015. “PHEnix.” 2015. https://github.com/phe-bioinformatics/PHEnix. Accessed 31/08/2019. Web Archive: NA.

Pruimboom-Brees, I. M., T. W. Morgan, M. R. Ackermann, E. D. Nystrom, J. E. Samuel, N. A. Cornick, and H. W. Moon. 2000. “Cattle Lack Vascular Receptors for Escherichia Coli O157:H7 Shiga Toxins.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (19): 10325–29. https://doi.org/10.1073/pnas.190329997.

Quick, J., N. D. Grubaugh, S.T. Pullan, I. M. Claro, A. D. Smith, K. Gangavarapu, G. Oliveira, et al. 2017. “Multiplex PCR Method for MinION and Illumina Sequencing of Zika and Other Virus Genomes Directly from Clinical Samples.” *Nature Protocols* 12 (May): 1261. https://doi.org/10.1038/nprot.2017.066.

Quick, J., N. J. Loman, S. Duraffour, J. T. Simpson, E. Severi, L. Cowley, J. A. Bore, et al. 2016. “Real-Time, Portable Genome Sequencing for Ebola Surveillance.” *Nature* 530 (February): 228. https://doi.org/10.1038/nature16996.

R Core Team. 2019. “R: A Language and Environment for Statistical Computing.” 2019. https://www.r-project.org/. Accessed: 01/09/2019. Web Archive: https://web.archive.org/web/20190831232313/http://www.r-project.org/.

Raeside, C., J. Gaffé, D. E. Deatherage, O. Tenaillon, A. M. Briska, R. N. Ptashkin, S. Cruveiller, et al. 2014. “Large Chromosomal Rearrangements during a Long-Term Evolution Experiment with Escherichia Coli.” Edited by Søren Baquero Molin Fernando. *MBio* 5 (5): e01377-14. https://doi.org/10.1128/mBio.01377-14.

Ratnam, S., S. B. March, R. Ahmed, and G.S. Bezanson. 1988. “Characterization of

Escherichia Coli Serotype 0157 : H7." *Journal of Clinical Microbiology* 26 (10): 2006–12.

Ravindran, P., and A. Gupta. 2015. "Image Processing for Optical Mapping." *GigaScience* 4 (November): 57. https://doi.org/10.1186/s13742-015-0096-z.

Reid, S. D., C. J. Herbelin, A. C. Bumbaugh, R. K. Selander, and T. S. Whittam. 2000. "Parallel Evolution of Virulence in Pathogenic Escherichia Coli." *Nature* 406 (6791): 64–67. https://doi.org/10.1038/35017546.

Ribot, E. M., M. A. Fair, R. Gautom, D. N. Cameron, S. B. Hunter, B. Swaminathan, and T. J. Barrett. 2006. "Standardization of Pulsed-Field Gel Electrophoresis Protocols for the Subtyping of Escherichia Coli O157:H7, Salmonella, and Shigella for PulseNet." *Foodborne Pathogens and Disease* 3 (1): 59–67. https://doi.org/10.1089/fpd.2006.3.59.

Rocha, E. P. C. 2004. "The Replication-Related Organization of Bacterial Genomes." *Microbiology* 150 (6): 1609–27. https://doi.org/10.1099/mic.0.26974-0.

Saile, N., A. Voigt, S. Kessler, T. Stressler, J. Klumpp, L. Fischer, and H. Schmidt. 2016. "Escherichia Coli O157:H7 Strain EDL933 Harbors Multiple Functional Prophage-Associated Genes Necessary for the Utilization of 5-N-Acetyl-9-O-Acetyl Neuraminic Acid as a Growth Substrate." *Applied and Environmental Microbiology* 82 (19): 5940–50. https://doi.org/10.1128/AEM.01671-16.

Scheutz, F., L. D. Teel, L. Beutin, D. Pierard, G. Buvens, H. Karch, A. Mellmann, et al. 2012. "Multicenter Evaluation of a Sequence-Based Protocol for Subtyping Shiga Toxins and Standardizing Stx Nomenclature." *J Clin Microbiol* 50 (9): 2951–63. https://doi.org/10.1128/JCM.00860-12.

Schmidt, H., and M. Hensel. 2004. "Pathogenicity Islands in Bacterial Pathogenesis." *Clinical Microbiology Reviews* 17 (1): 14–56. https://doi.org/10.1128/cmr.17.1.14-56.2004.

Schmidt, S., A. Gerasimova, F. A. Kondrashov, I. A. Adzuhbei, A. S. Kondrashov, and S. Sunyaev. 2008. "Hypermutable Non-Synonymous Sites Are under Stronger Negative Selection." *PLOS Genetics* 4 (11): e1000281.

https://doi.org/10.1371/journal.pgen.1000281.

Schutz, K., L. A. Cowley, S. Shaaban, A. Carroll, E. McNamara, D. L. Gally, G. Godbole, C. Jenkins, and T. J. Dallman. 2017. “Evolutionary Context of Non-Sorbitol-Fermenting Shiga Toxin-Producing Escherichia Coli O55:H7.” *Emerging Infectious Diseases* 23 (12): 1966–73. https://doi.org/10.3201/eid2312.170628.

Seemann, T. 2014. “Prokka: Rapid Prokaryotic Genome Annotation.” *Bioinformatics* 30 (14): 2068–69. https://doi.org/10.1093/bioinformatics/btu153.

Shaaban, S., L. A. Cowley, S. P. McAteer, C. Jenkins, T. J. Dallman, J. L. Bono, and D. L. Gally. 2016. “Evolution of a Zoonotic Pathogen: Investigating Prophage Diversity in Enterohaemorrhagic Escherichia Coli O157 by Long-Read Sequencing.” *Microbial Genomics* 2 (12). https://doi.org/10.1099/mgen.0.000096.

Sharp, P. M., D. C. Shields, K. H. Wolfe, and W. H. Li. 1989. “Chromosomal Location and Evolutionary Rate Variation in Enterobacterial Genes.” *Science* 246 (4931): 808 LP – 810. https://doi.org/10.1126/science.2683084.

Smith, T. F., and M. S. Waterman. 1981. “Identification of Common Molecular Subsequences.” *Journal of Molecular Biology* 147 (1): 195–97. https://doi.org/10.1016/0022-2836(81)90087-5.

Stewart, A. C., B. Osborne, and T. D. Read. 2009. “DIYA: A Bacterial Annotation Pipeline for Any Genomics Lab.” *Bioinformatics* 25 (7): 962–63. https://doi.org/10.1093/bioinformatics/btp097.

Stothard, P. 2000. “The Sequence Manipulation Suite: JavaScript Programs for Analyzing and Formatting Protein and DNA Sequences.” *BioTechniques* 28 (6): 1102–4. https://doi.org/10.2144/00286ir01.

Sullivan, M. J., N. K. Petty, and S. A. Beatson. 2011. “Easyfig: A Genome Comparison Visualiser.” *Bioinformatics* 27 (7): 1009–10. https://doi.org/10.1093/bioinformatics/btr039.

Tobe, T., S. A. Beatson, H. Taniguchi, H. Abe, C. M. Bailey, A. Fivian, R. Younis,

et al. 2006. “An Extensive Repertoire of Type III Secretion Effectors in Escherichia Coli O157 and the Role of Lambdoid Phages in Their Dissemination.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (40): 14941–46.

Underwood, A. P., T. Dallman, N. R. Thomson, M. Williams, K. Harker, N. Perry, B. Adak, et al. 2013. “Public Health Value of Next-Generation DNA Sequencing of Enterohemorrhagic Escherichia Coli Isolates from an Outbreak.” *Journal of Clinical Microbiology* 51 (1): 232–37. https://doi.org/10.1128/JCM.01696-12.

Wickham, H. 2009. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.

Xiao, J., Z. Zhang, J. Wu, and J. Yu. 2015. “A Brief Review of Software Tools for Pangenomics.” *Genomics, Proteomics & Bioinformatics* 13 (1): 73–76. https://doi.org/10.1016/j.gpb.2015.01.007.

Xu, X., S. P. McAteer, J.J. Tree, D. J. Shaw, E. B. K. Wolfson, S. A. Beatson, A. J. Roe, et al. 2012. “Lysogeny with Shiga Toxin 2-Encoding Bacteriophages Represses Type III Secretion in Enterohemorrhagic Escherichia Coli.” *PLoS Pathogens* 8 (5): e1002672. https://doi.org/10.1371/journal.ppat.1002672.

Zhou, Y., Y. Liang, K. H. Lynch, J. J. Dennis, and D. S. Wishart. 2011. “PHAST: A Fast Phage Search Tool.” *Nucleic Acids Research* 39 (SUPPL. 2): 1–6. https://doi.org/10.1093/nar/gkr485.