THE UNIVERSITY
of EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

The use of genome sequencing to investigate the molecular basis of bacteria-phage interaction of the *Escherichia coli* O157 typing phages and the elucidation of the biological and public health significance of phage type

By

Lauren A Cowley

University of Edinburgh

Public Health England

Submitted for examination in the degree of

Doctor of Philosophy

July 2016

Signed Declaration

I, Lauren A Cowley, confirm that the work presented in this thesis was composed by me and is my own. I confirm that the work has not been submitted for any other degree or professional qualification. Where information has been derived from other sources, I confirm that this has been indicated in the thesis. The included publication in section 7.2 is my own work.

Lauren A Cowley

July 2016

Table of Contents

List of Figures

CHAPTER 4: IMPACT OF SHORT AND LONG READ SEQUENCING TECHNOLOGIES IN UNDERSTANDING PT DIFFERENCES 108

CHAPTER 5: THE USE OF TRADIS TO ELUCIDATE GENES INVOLVED IN TYPING PHAGE SUSCEPTIBILITY 136

List of Tables

Acknowledgements

My PhD supervisors Claire Jenkins, Tim Dallman and David Gally for their guidance throughout my PhD.

My family for their love and support during my thesis writing and my friends for providing me with endless amusement on social occasions in between writing.

David Bowie for his musical genius (Suffragette city is the unofficial soundtrack to this thesis) and the Chandos pub for always providing a cold one when it is needed.

Layman's Summary

Bacteriophages are viruses that infect bacteria but they can also be used for typing of bacteria based on the bacteria's susceptibility and resistance profiles. Typing of bacteria is very important for epidemiological and clinical uses and it helps determine whether a strain is particularly virulent or related to other cases i.e. part of an outbreak. To understand bacterial virulence and outbreaks better we need to understand the molecular mechanisms of phage typing to determine why we see certain phage types more commonly or associated with more serious outbreaks than others. This study has focused on the phage typing scheme of Shiga Toxin producing *Escherichia coli* (STEC) O157, a cause of severe gastrointestinal symptoms. This phage typing scheme has been used in the UK to type STEC O157 for the last 25 years.

This thesis details the use of whole genome sequencing to understand the genetic basis for the phage susceptibility and resistance observed in the phage typing scheme. The main findings are the identification of several different genetic factors that influence phage type.

The typing phages were classified into 4 groups based on their genetic similarity that correlated well with their infectivity. This has helped in the subsequent analysis of resistance and susceptibility of the host. Initial analysis showed that phage type was often determined by the accessory genome and that short read sequencing was insufficient to resolve this. The use of long read sequencing revealed that single genetic events such as acquisition of a plasmid or bacteriophage can change phage type. Transposon directed insertion site sequencing was used as a genome wide mutagenesis approach that showed that phage infectivity was controlled by a number of different genetic mechanisms including the stringent starvation protein A and the Sap operon but that the interaction of the genetic components involved was complex. This study has shown that, although complex, genetic determinants for PT can be mined from the genome and allow us to understand the evolution of this zoonotic pathogen between host species and during outbreaks.

Abstract

Background

Shiga toxin producing *Escherichia coli* (STEC) O157 causes severe gastrointestinal disease and haemolytic uremic syndrome, and has a major impact on public health worldwide with regular outbreaks and sporadic infection. Phage typing, i.e. the susceptibility of STEC O157 strains to a bank of 16 bacteriophages, has been used in the UK to differentiate STEC O157 for the past 25 years and the phage type (PT) can be an epidemiological marker of strains associated with severe disease or associated with cases that occur from foreign travel. However, little is known about the molecular interactions between the typing phages (TP) and STEC O157. The aims of this thesis were to use whole genome sequencing to elucidate the genetic basis for phage typing of STEC O157 and through this understand genetic differences between strains relevant to disease severity and epidemiology.

Results

Sequencing the STEC O157 TPs revealed that they were clustered into 4 groups based on sequence similarity that corresponded with their infectivity. Long read sequencing revealed microevolutionary events occuring in STEC O157 genomes over a short time period (approximately 1 year), evidenced by the loss and gain of prophage regions and plasmids. An IncHI2 plasmid was found responsible for a change in Phage Type (PT) from PT8 to PT54 during two related outbreaks at the same restaurant. These changes resulted in a strain (PT54) that was fitter under certain growth conditions and associated with a much larger outbreak (140 as opposed to 4 cases). TraDIS (Transposon directed Insertion site sequencing) was used to identify 114 genes associated with phage sensitivity and 44 genes involved in phage resistance, emphasising the complex nature of identifying specific genetic markers of phage susceptibility or resistance. Further work is required to prove their phage-related functions but several are likely to encode novel phage receptors. Deletion of a Stx2a prophage from a PT21/28 strain led to a strain that typed as PT32, supporting the concept that the highly pathogenic PT21/28 lineage I strains emerged from Stx2c+ PT32 strains in the last two decades by acquisition of Stx2a-encoding prophages.

Conclusions

This body of work has highlighted the complexity of bacteriophage interaction and investigated the genetic basis for susceptibility and resistance in *E. coli*. The grouping of the TPs showed that resistance or susceptibility to all members of a typing group was likely to be caused by one mechanism. IncHI2 was identified as one of the markers for the PT54 phenotype. The Stx2a prophage region was associated with the switch from PT32 to PT21/28, although PT32 strains containing both Stx2a and Stx2c-encoding prophages have been isolated and can provide insights into phage variation underpinning the susceptibility to the relevant typing phages. The TraDIS results indicated that susceptibility or resistance was governed by multiple genetic factors and not controlled by a single gene. The significance of LPS for initial protection from phage adsorption was evident and a number of novel genes controlling phage susceptibility and resistance identified including the Sap operon and stringent starvation protein A respectively. While SNP-based typing provides an excellent indication of the evolution and relatedness of strains, phage typing can provide real insights into short term evolution of the bacteria as PTs can be altered by mobile elements such as prophages and plasmids. This study has shown that, although complex, genetic determinants for PT can be mined from the genome and allow us to understand the evolution of this zoonotic pathogen between host species and during outbreaks.

Chapter 1: Introduction

1.1 Escherichia coli

The bacterial species *Escherichia coli* is one of the most versatile, widely exploited by humans and commonly found on the planet. Humans are colonised with *E. coli* on the day of their birth and coexist in a synergistic relationship. Most strains of *E. coli* are harmless to humans and may even play a synergistic role [1]. However, it has often been observed that bacteria can move from commensal to opportunistic to pathogenic through the steady acquisition of virulence factors and there are types of *E. coli* that can be pathogenic to humans. These include Extraintestinal Pathogenic *E. coli* (ExPEC) that can cause sepsis, UTI, abscesses and meningitis, and a wide variety of diarrheagenic *E. coli* (DEC). The types of DEC are known as enteropathogenic *E. coli* (EPEC), enterotoxigenic *E. coli* (ETEC), enteroaggregative *E. coli* (EAEC), enteroinvasive *E. coli* (EIEC), diffusely adherent *E. coli* (DEAC) and enterohaemorrhagic *E. coli* (EHEC) [2].

Enterohaemorrhagic *E. coli* are a subset of Shiga toxin producing *E. coli* (STEC) with Shiga toxins encoded on lambdoid bacteriophages integrated into the bacterial genome. Any *E. coli* isolate encoding a Shiga toxin (Stx) variant can be termed an STEC, while classical EHEC strains were originally defined as STEC that encode a type III secretion system (specifically the Locus of Enterocyte Effacement (LEE)) and are associated with human disease, typically bloody diarrhoea. The prophage encoded *stxAB* operon is associated with the host expression of Shiga toxin 1 (Stx1) and Shiga toxin 2 (Stx2) subtypes. STEC O157 is the most prevalent serogroup responsible for causing gastrointestinal disease in the UK. Of the DECs, STEC has the most severe impact on public health in the UK causing severe bloody diarrhoea and haemolytic uraemic syndrome (HUS) in 5-10% of cases. The main reservoir of STEC O157 are ruminants[3] so contaminated meat, vegetables or water and contact with cattle, sheep or goats or their environments are often associated with outbreaks. Person-to-person transmission also occurs in closed environments, such as nurseries and households. The infectious dose for STEC O157 is considered low at <100 cells of bacteria [4], making it a highly dangerous pathogen.

1.2 Evolution and genome arrangement of STEC O157

Originally it was proposed that the evolutionary steps of the acquisition of the Shiga toxins were firstly Shiga toxin 2 (Stx2) and then Shiga toxin 1 (Stx1),followed by loss of the ability to ferment sorbitol in an ancestoral strain related to *E. coli* O55:H7 [5] . Feng et al delineated the population structure of STEC O157 into three lineages I, I/II and II. These lineages have also been found in the phylogenetic analysis of the UK strains ([6]) but the Dallman study performed a timed phylogeny that revealed that the ancestor was a *stx2c+* *E. coli* O55 that lost the ability to ferment sorbitol followed by multiple acquisitions of *stx1* (fig. 1.1). Figure 1.1 shows the evolutionary model proposed by Dallman et al based on a timed phylogeny of over 1000 genomes. Not only were they able to observe the gain of new Shiga phages but also the loss and replacement of the *stx2c* phage in different sub-lineages of lineage I.

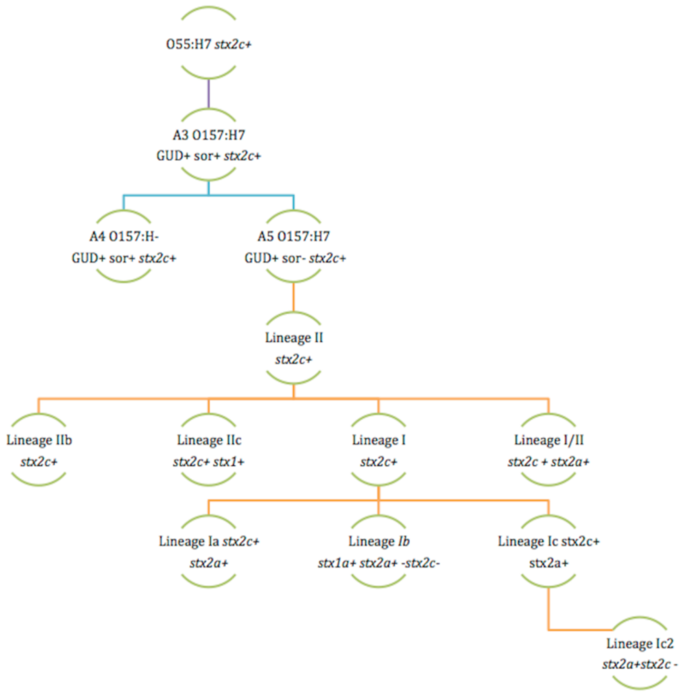


Figure 1.1 Evolutionary model of STEC stx phage gain taken from Dallman et al. 2015 [6]. Revealing the step wise acquisition of stx phage in multiple acquisitions and in multiple lineages as well as the loss of the ability to ferment sorbitol from an STEC O55 ancestor.

An exemplar strain (termed the Sakai strain) associated with an outbreak of STEC O157 in Japan was sequenced in 2001 and revealed that STEC O157 has a 5.5mb genome, with 5361 protein coding regions and 50.5% GC content (fig. 1.2). The STEC O157 chromosome is 1.4mb larger than K-12 MG1655 and there is 4.1mb conserved between them. The Sakai chromosome revealed 18 prophage regions,was the first evidence of the high rate of phage lysogeny within the STEC O157 clone and revealed the mosaic nature of the genome [7].

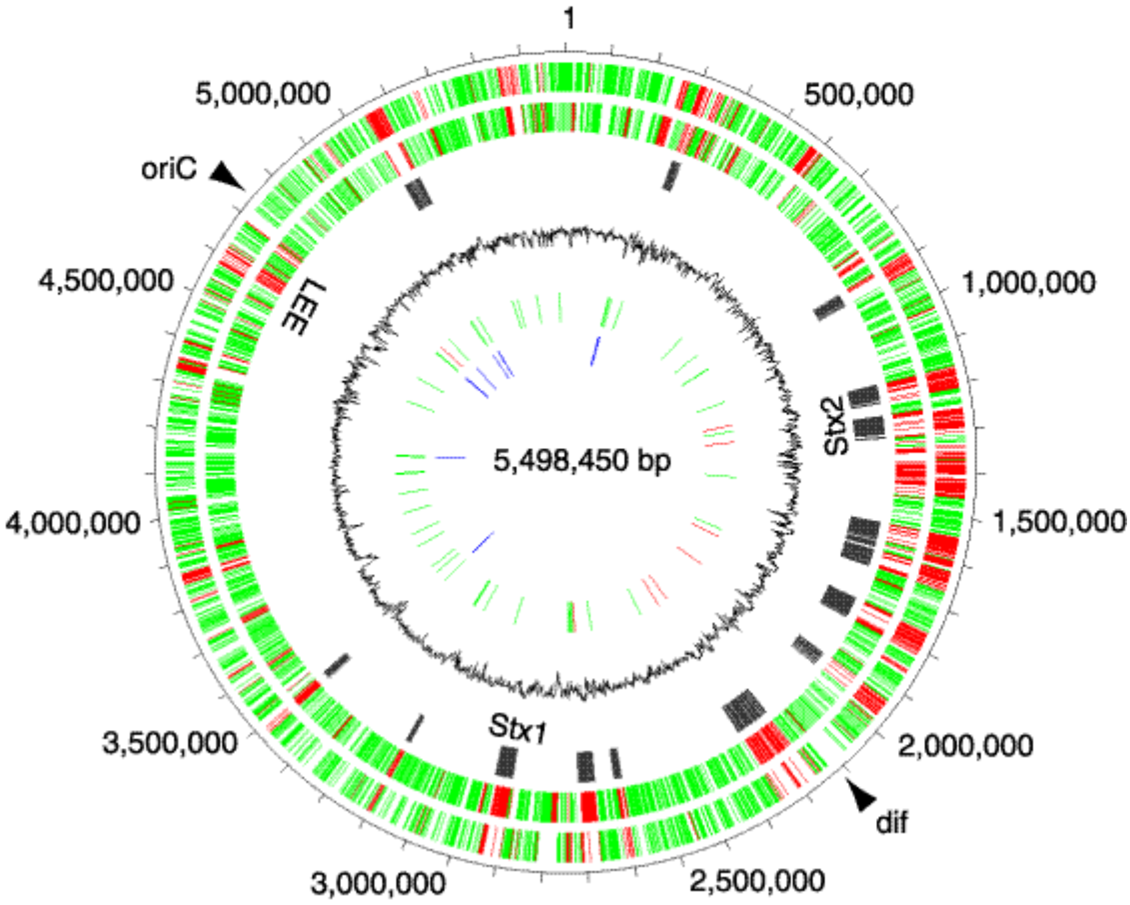


Figure 1.2 Genomic arrangement of STEC O157 taken from [7]. The outermost circle indicates the chromosomal location in base pairs (each tick is 100Kb). The second and the third show predicted ORFs transcribed in the clockwise and anticlockwise directions, respectively. ORFs conserved in *E.coli* K-12 (MG1655) are depicted in green and those not present in K-12 in red. The fourth circle shows the locations of ORFs on prophage genomes in grey . The fifth shows the 20Kb window-average of G+C percent in relation to the mean value of the chromosome. The locations of tRNA and rRNA (blue) genes are shown in the sixth and seventh circles, respectively. tRNAs conserved in K-12 are depicted in green, and those absent in K-12 are in red.

1.2 Pathogenicity factors of STEC O157

Stx2 has 55-60% identity in DNA to Stx1 so is immunologically distinct. There are several subtypes of *stx2*, with Stx2a and/or Stx2c most commonly found in STEC O157. There a four subtypes of *stx1* (a-d) and seven subtypes of *stx2* (a-g). In a study of 349 strains of STEC O157 in England 236 had multiple copies of different Stx subtypes with 33 of these being variants between the subtype alleles and 21 with identical subtype copies [8]. The subtypes are generally differentiated by PCR [9]. Strains that have Stx2a are significantly associated with the causation of bloody diarrhoea and HUS [8]. In humans Stx bind to glycosphingolipid globotriaosylceramide (Gb3/CD77), a cell-surface receptor, are internalised by clathrin-dependent endocytosis and depurinate 28S eukaryotic rRNA resulting in inhibition of protein synthesis; eventually leading to cytokine release and apoptotic cell death [10]. Gb3 receptors are found in high abundance on the surface of endothelial cells within that line the vasculature of the kidneys, heart and brain, and this is one reason for the association of STEC with kidney failure (HUS), heart disorders and neurological disorders [11]. Cattle endothelial cells lack Gb3 receptors so are not susceptible to the same pathological consequences of Stx exposure as observed in humans [12]. Shiga-toxigenic bacteriophage have have the ability to lysogenise a wide range of *E. coli* and *Shigella* and therefore carry the potential to transfer the *stx* gene in clinical and agricultural environments resulting in widespread dissemination [13].

As well as the Shiga toxins, STEC O157 has other genetic elements that are important for their pathogenicity. The locus of enterocyte effacement (LEE) encodes a type III secretion systems (T3SS) responsible for bacterial attachment to the intestinal epithelium and destruction of the microvilli brush border in the human gut surrounding the point of attachment. A site of bacterial attachment is created by rearrangement of actin filaments within host enterocytes and sometimes results in the formation of an actin-dense pedestal. This process is driven by a cascade of effector proteins translocated into the host cell [14, 15] by the T3SS. EHEC and EPEC produce a secretion needle extension known as a translocon which is composed of EspA [16] and this links through to a pore formed in the host cell membrane composed of EspD and EspB[17]. There are over 40 different secreted effector proteins exported by STEC O157 strains. One of the first to be characterised was the translocated intimin receptor - Tir [18] which was shown to act as a receptor for the bacterial outer membrane protein, intimin. Their interaction triggers a signalling cascade that leads to cytoskeleton rearrangements resulting in intimate attachment.

Stx2a and Stx2c have been shown to repress T3S and stx2a is most commonly found in the strains associated with supershedding from cattle. The repression of T3S is achieved by repression of Ler induction of LEE1, it is thought that this repression has evolved to ensure that the phages have control of this critical colonization factor to induce co-dependence on them for colonisation [19]. Regulators that control the GAD acid stress response also indirectly control T3S, they are encoded on prophages and usurp the conserved GAD acid stess resistance system to regulate T3S by increasing the expression of GadE (YhiE) and YhiF following attachment to bovine epithelial cells [20]. The LEE is also known to be regulated by bacterial nucleoid associated proteins (NAPs) [21] with the histone-like nucleoid structuring protein (H-NS) repressing the LEE under non-inducing conditions such as low temperature.

1.3 Impact on Public Health

STEC O157 symptoms can range from mild gastroenteritis to severe bloody diarrhoea and HUS and thrombotic thrombocytopenic purpura (TTP). The very young, elderly and immune-compromised are at particular risk of HUS and TTP [22]. During a recent Public Health England (PHE) study incidence was found to be 1.8 per 100000 person-years with up to 34.3% of cases being hospitalised [23]. The Gastrointestinal Bacterial Reference Unit at PHE receives approximately 1000 STEC O157 samples per year. Outbreaks are often linked with food poisoning, i.e. restaurant or food supply associated, or the farming environment. STEC O157 can pose a large threat to public health and a recent outbreak in the UK in 2009 resulted in 93 cases of infection and ~22% developing HUS[24].

The source-sink model is used to describe the STEC O157 transmission route, with humans being the sink (generally no further progression in terms of survival of the organism) and ruminants being the source[10]. In the source, the microbe colonises the rectum of cattle and is shed. Cattle can sometimes be categorized as 'supershedders' and can shed up to 10^7 STEC O157 per gram of faeces[25]. These 'supershedders' have usually been colonized at a lymphoid follicle-dense mucosal region at a short distance proximal to the recto-anal junction [26, 27]. In the sink, human infections usually result in 5-10% of patients developing HUS (varying significantly with age, sex and *stx* subtype), commonly 3-8 days after the onset of infection [23, 28].

Incidence of STEC is higher in children aged 1-4 (7.63 per 100000 person-years), females and white ethnic groups with progression to HUS most frequent in females and children [23]. It was also found that incidence of STEC was four times higher in people living in rural areas than urban ones due to exposure to livestock and the risk associated with that for sporadic STEC infection [23]. STEC infection can result in secondary transmission in certain settings such as schools or nurseries

[29]. School outbreaks account for a fifth of all the STEC cases reported in England and Wales annually [30]. An enhanced understanding of the pathogen and why it is associated with certain risk factors could improve mitigation strategies to prevent outbreaks.

1.4 Sequencing technologies and the use of bioinformatics in epidemiology and public health

DNA sequencing was first described by Sanger in 1977 [31] and has developed rapidly in recent years with the introduction of next-generation sequencing technologies. This began with the introduction of high-throughput sequencing-by-synthesis technology developed by 454 Life Sciences in 2005 [32]. Whole genome sequencing for the public health surveillance, detection and investigation of outbreaks of the major foodborne bacterial pathogens has been adopted by PHE [33-35]. The technology that has become the main workhorse for routine sequencing both for Eukaryotic and Prokaryotic use is the Illumina platform. These technologies provide short read sequencing to high depth and this data can be used to infer gene presence or absence, Multi Locus Sequence Typing (MLST), species identification and phylogenetic typing [36]. Phylogenetic typing allows isolates to be clustered based on relatedness in the core (shared) genome by the number of single nucleotide polymorphisms (SNPs) that they share. Hierarchical clustering of SNPs has proved very useful in outbreak investigations and it has been shown that isolates of STEC O157 with <5 SNPs between them are likely to be epidemiologically linked [37].

More recently there has been a focus on the development and improvements of long-read sequencing as it is highly useful for assembling complex genomes with highly repetitive elements like prophages. Technologies such as PacBio and MinION have led the long-read sequencing market. PacBio so far has provided the more accurate and reliable data with several genomes being finished into single contiguous sequences [38]. However MinION provides an advantage over PacBio technology with its portability. The MinION is a palm-sized nanopore sequencing device that has already proved its applicability in its use during a high profile outbreak in the field setting [39].

1.5 Detection and typing

Unlike ~90% of *E. coli* in the human gut, STEC O157 are resistant to cefixime and tellurite and do not ferment sorbitol so can be identified by culturing faecal specimens on selective indicator media, such as cefixime tellurite sorbitol MacConkey, on which colourless (translucent) colonies grow following overnight incubation at 37°C. Identification can be confirmed by agglutinating the colonies with specific antiserum. Strains have historically been differentiated by phage typing or pulsed-field gel electrophoresis and multilocus variable number tandem repeat analysis [40]. In 2015, whole genome sequencing(WGS) and SNP-based typing was implemented at PHE. WGS provides an indication of evolutionary associations and is based on DNA sequences, WGS data is also used to mine for genotyping data such as serotype, MLST, kmer based identification and virulence genes [37].

Phage-typing has been used to categorize *E. coli* outbreaks and sporadic cases but little is known about the molecular basis for the interaction between typing phages and the bacterial phage types (PT). There are differences in the epidemiology and virulence associated with specific phage types and understanding the genetic differences between PTs could define determinants required for the occupation of specific environmental niches and disease severity.

In the past, epidemiologists have used PT in their case definition for outbreaks, for example, during a recent school outbreak of PT32 STEC O157 in Staffordshire [29]. However, PT provides low level strain discrimination. In a recent study evaluating the use of WGS data [33] an outbreak of PT2 STEC O157 associated with watercress consumption was ultimately found to be two separate but concurrent outbreaks associated with different sources of contamination of the watercress. In certain cases PT can confound epidemiological investigations that may link unrelated cases because they share the same PT.

1.6 Phage typing scheme of STEC O157

Phage-typing of STEC O157 is a scheme based on the use of 16 bacteriophages that produce a phage infection profile for a strain scored on the level of lysis achieved by each phage [41]. Certain PTs are more likely to be associated with human infection and so far there is little understanding of the basis for this. Further insight into relevant strain differences can be gained by sequencing different bacterial phage types and by sequencing the typing phages themselves and defining the basis of their infection selectivity.

Phage-typing has been used for decades as a definitive phenotypic method for epidemiological typing. From the 1920s to the 1970s, phage typing schemes for a range of bacteria were devised as another form of typing to discriminate and provide a further level of discrimination to enhance serotyping. As well as the STEC O157 phage typing scheme, there are known phage typing schemes for *Salmonella enterica serovar Enteritidis, Salmonella typhimurium, Vibrio cholerae, Staphylococcus aureus, Campylobacter, Clostridium difficile, Corynebacterium* and many more. Recently *Mycobacterium tuberculosis* phage typing was replaced by whole genome sequencing [42]. However some phage typing schemes are still used widely [43, 44].

The 16 phages in the STEC phage-typing scheme are made up of 14 T4 phages (fig. 1.3) and 2 T7 phages (fig. 1.4). An example of a T7 phage has been sequenced previously and T7 are known to consist of a single ‘chromosome’ carrying about 30 genes [45]. The 5’ end genes of the chromosome are expressed at an early stage of infection and their products are involved in the induction of host RNA polymerase for transcription and control the expression of other phage genes in a positive feedback mechanism. Genes that are expressed later are involved in the metabolism of phage DNA and code for capsid proteins or are involved in the assembly of infective progeny particles [45]. T4 phages have much larger genomes with 300 putative genes, only 62 of these have been found to be ‘essential’ under laboratory conditions [46]. The order of expression works in a similar way to T7 phage.

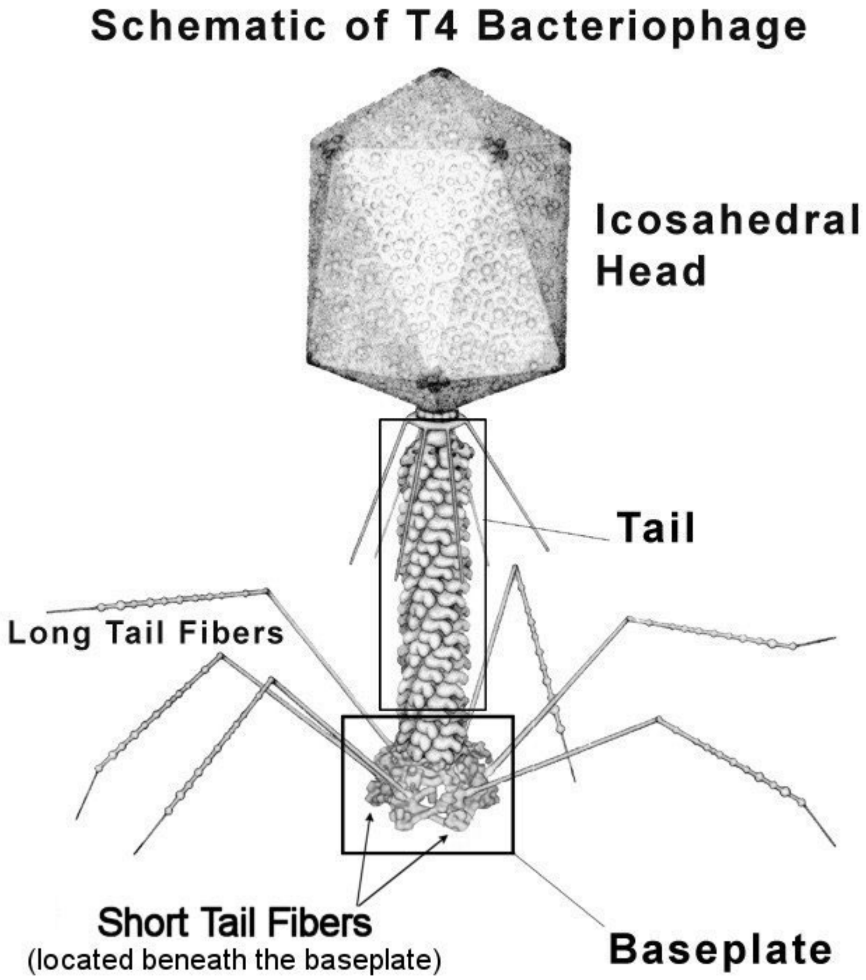


Figure 1.3 Schematic diagram of a T4 bacteriophage showing the Icosahedral (20 sided) head containing the DNA, the tail fibers used to recognise the host and anchor the phage, the Baseplate which is the site of adherence to the bacterial cell and the tail which is used to insert the DNA into the cell. Figure credit to Petr Leiman (Purdue University) adapted from a drawing by Fred Eiserling (UCLA).

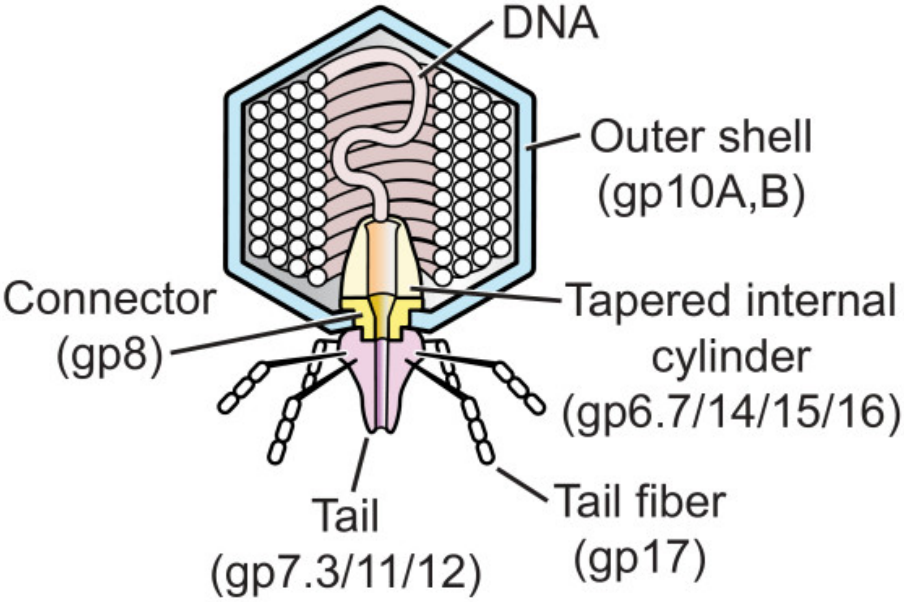


Figure 1.4 Schematic diagram of a T7 bacteriophage showing the outer shell or head containing the DNA, the tail which is characteristically much shorter than that seen on T4, the host recognition and anchoring tail fibers and the connector and tapered internal cylinder which are used to insert the DNA. The phage gene proteins that encode these elements are listed underneath each [47].

Little is known about the molecular basis for the interaction between phages and strains of different PT, however we can interrogate the phage infection profile of who-infects-whom as a bipartite (two-mode) network. Two common methods for analysing community structure in bipartite data are nestedness and modularity. Nestedness is a way of measuring the ranges of both host resistance and phage infectivity across a specialist to generalist gradient. Specialists are assumed to have strategies that are subsets of those which are more generalised. Modularity is the degree to which a network can be split into distinct modular groupings of phage and bacteria such that there are more infections within rather than between groups [48].

1.7 Epidemiology and link to PT

Sixty-one percent of recently reported (PHE in-house data) STEC O157 cases are PT8 or PT21/28 so these PTs may be more prevalent in animals or may be more likely to be associated with more severe disease. PT8 is the PT most commonly associated with foreign travel and PT21/28 is more commonly associated with domestic cases. There is generally more diversity among foreign travel associated cases than domestic cases indicating that there are a few dominant clones circulating in the UK and more diversity sampled in the rest of the world. Other common PTs in the UK are 2, 4 and 32.

Historically PT2 was more dominant than PT21/28 in the UK but there has been a shift in prevalence over time [49]. PT2 accounted for 45.5% of cases in 1994 but declined to just 16% in 2002 and 3.4% in 2012 (fig. 1.5). Conversely, PT21/28 accounted for 9% in 1994, 33.1% in 2002 and 29.6% in 2012. PT8 has also increased in occurrence from 1994 to 2012. Examples of differences in dominant PTs could be the result of clone evolution and the change in its susceptibility/resistance to the typing phages or could be an example of a new strain replacing the dominant one. With regards to the replacement of PT21/28 with PT2, it has been shown that this is the result of strain replacement [6].

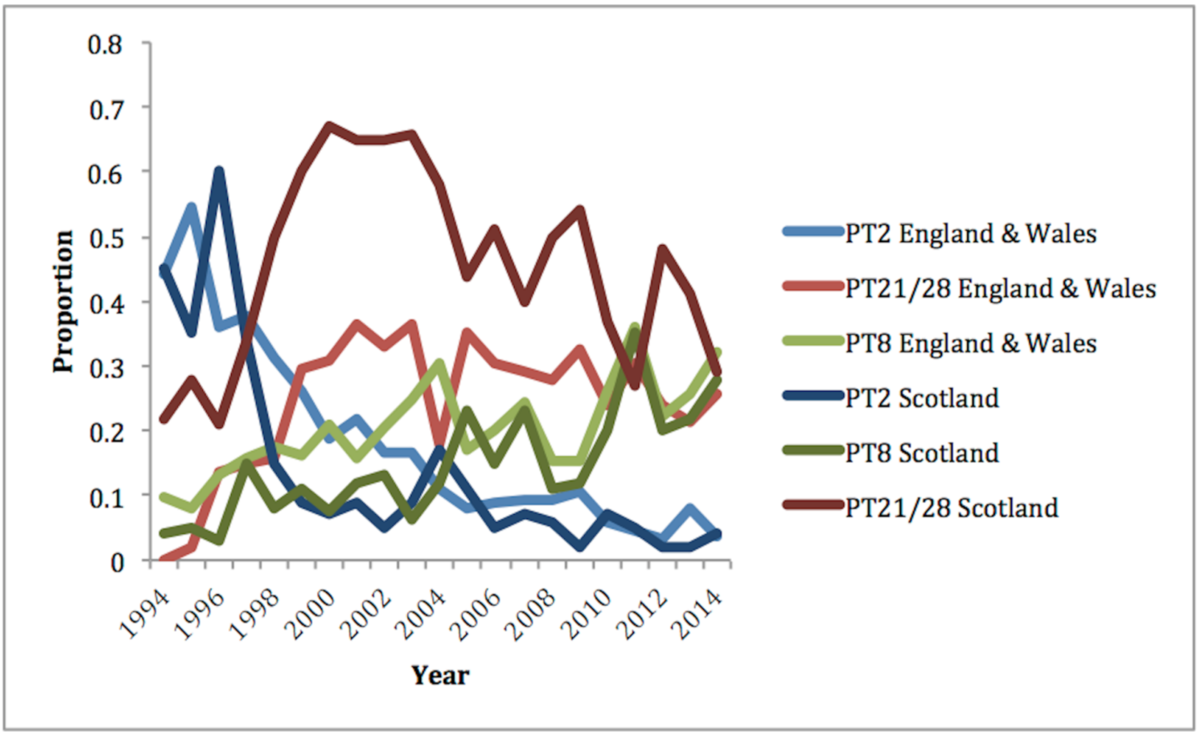


Figure 1.5 Proportion of cases of the predominant phage types in England and Wales, and Scotland over the last 20 years.

Figure 1.6 illustrates that genetically related clades of STEC strains are often dominated by a single PT but will include closely related strains that have evolved to become different PTs through an unknown mechanism. This can be observed in the PT21/28 clade with sporadic incidences of PT32, PT33 and PT4 occurring within it. This is also seen, although less commonly, in the large PT8 clade with a few incidences of PT 31, PT 54, PT 14 and PT 1. This is evidence for the transient nature of some phage resistance/susceptibility and the likelihood of PT being linked to mobile elements such as plasmids or prophages. However it is also clear from figure 1.6 that PT can be conserved for 1000s of generations of STEC and therefore does provide an indication of relatedness.

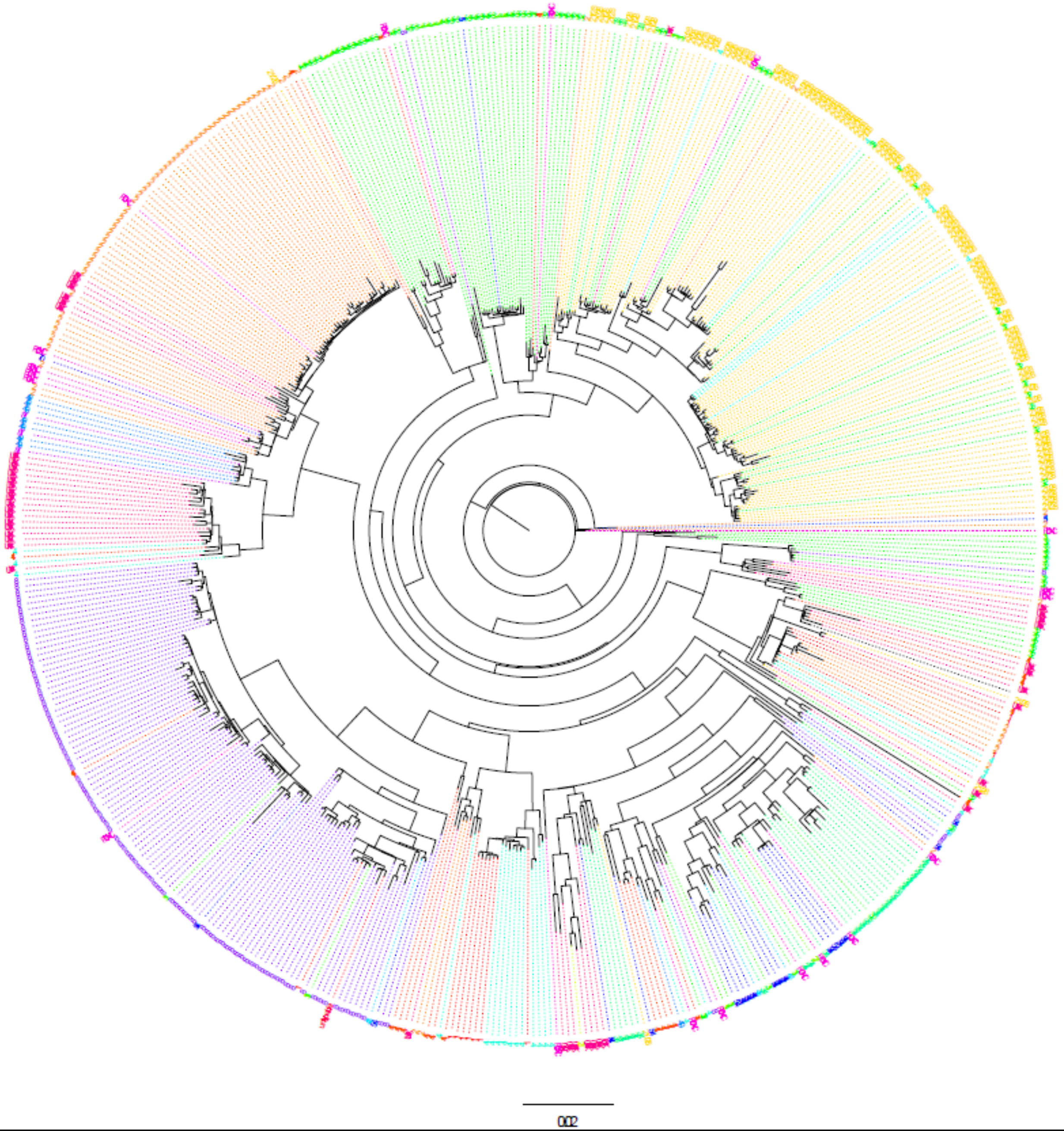
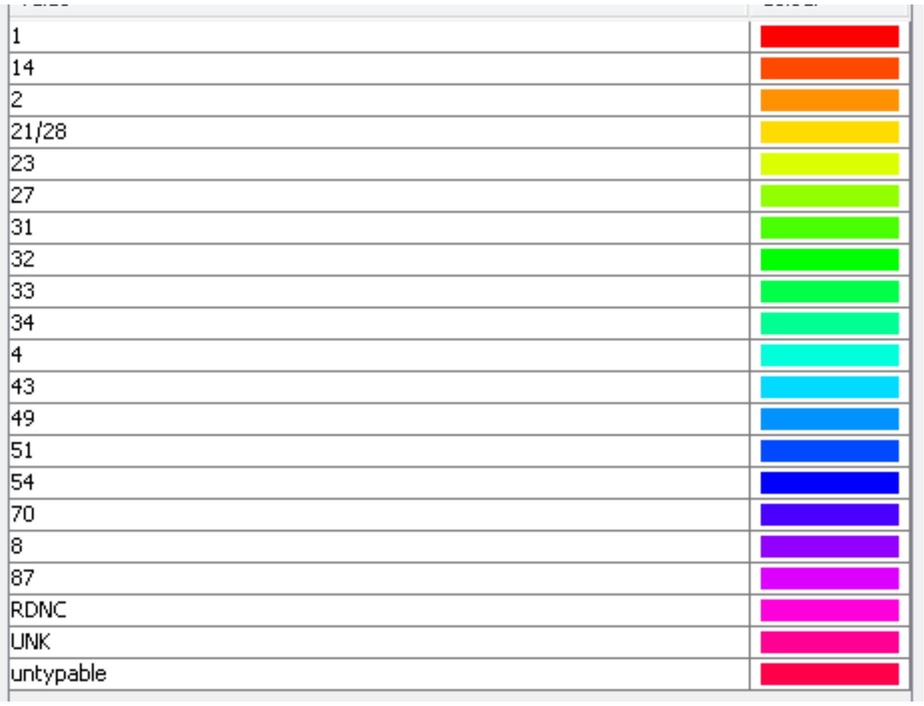


Figure 1.6 Maximum likelihood phylogenetic tree of 544 representative cases of STEC of various PT from the last 20 years in the UK. PT annotated colour coded as below:



1	🟥
14	🟧
2	🟧
21/28	🟨
23	🟨
27	🟩
31	🟩
32	🟩
33	🟩
34	🟩
4	🟦
43	🟦
49	🟦
51	🟦
54	🟦
70	🟪
8	🟪
87	🟪
RDNC	🟪
UNK	🟥
untypable	🟥

1.8 Bacteriophage infection and prophages

Bacteriophages are viruses that infect bacteria and can cause bacterial lysis and cell death. There are a range of phages that infect *E. coli* and can enter a lytic or lysogenic phase after infection. During the lytic phase the phage causes cell lysis whereas during the lysogenic phase the phage becomes integrated into the host genome as a prophage. The prophage elements are often very important as they can encode key virulence factors so may provide a pathoadaptation to the strain. The first step in bacteriophage infection is attachment to cell surface receptors and phages of Gram-negative bacteria use a variety of cell-associated structures including pili, flagella, outer membrane proteins (Omp) or lipopolysaccharides (LPS) [50] to achieve this. Once the bacteriophage has attached to the bacterial cell it can inject its DNA into the cell through its tail. Bacteriophage also require host target genes and host encoded proteins for their replication; T7 bacteriophage is known to require host-encoded thioredoxin in *E. coli* for use as a subunit of its DNA polymerase [51]. Phages often have genetic switches that determine infection outcome and whether the lytic or lysogenic phase is entered. The switches contain three key genes: an integrase, a repressor and *cro* [52]. Phages can use insertion sites which are recognised by integrases, enzymes that mediate unidirectional site-specific recombination between two defined DNA sequences which also usually requires a bacterially-encoded integration host factor. Λ bacteriophage insertion requires *att*P which several integrase molecules bind to and that complex binds attB [53]. Commonly, phage are observed to have inserted into tRNA genes but this may be because these are the strains that have survived an insertion and tRNA insertion sites are the most evolutionarily fit. This may be because tRNA sites are often redundant in the genome so the cost of integration is low. T4 phage have endonuclease VII that supports robust DNA packaging by cleaving Holliday structures twice and DNA ligase seals the breaks to restore the original sequence to produce genetically sensible products to be packaged into phage heads [54]. T4 initiates replicaton from specialized structures on its genome called RNA-DNA hybrids at replication origins and the annealed RNA serves as a primer for leading-strand synthesis in one direction [55].

STEC O157 has a significant prophage content and the Sakai reference strain is known to contain 18 prophages and many have deletions and/or insertions in regions considered to be essential for phage reproduction. These are likely to be defective but can recombine and be induced to produce recombinant phages that may have a completely new set of features [56, 57]. Latent phage can be induced from their host bacterium by exposure to ultraviolet light and will subsequently lyse the bacterium and be produced from the cell [58]. Prophage regions of Sakai were shown to be inducible through Mitomycin C treatment in which prophages are de-repressed by a RecA-mediated mechanism to enter the lytic pathway [57]. Nonhomologous recombination can occur through the acquisition of 'morons' (units of more DNA), usually with a different G+C content and therefore indicating recent entry to the genome [59]. The shiga toxin genes are examples of morons as they are only expressed through transcription from a phage promoter during lytic growth after the prophage has been induced [60]. Very recent work on Stx phage type and Stx2 production levels has shown that stx2a can be further subtyped according to their replication proteins and certain subtypes of stx2a were shown to produce higher levels of shiga toxin than others [61]. This work further highlights the role of prophage variation in determining the production level of the virulence factors that phages encode.

1.9 Bacteriophage lysis mechanisms

It is known that bacteriophage lysis happens when phage-encoded lysozymes, a cell-wall destroying enzyme, accumulate during late protein synthesis. Lysozyme enzymes are synthesized when the coat proteins appear and result in the rupture of the cell wall. However, for the scheduling of lysis in T4 and T7 the accumulation of lysozyme is irrelevant [62]. Most phage genes that cause lysis have evolved to terminate the vegetative cycle by lysing the cell at precisely the right time. If lysis is caused too early then the progeny have not assembled, timing mechanisms operate both at the level of translation and at the level of protein function. This will make the gene(s) essential for plaque formation but not essential for accumulation of infective phage.

For T4 bacteriophage there are two described lysis mechanisms; lysis from within (LI) and lysis from without (LO). LI is dependent on the activity of phage-induced lysozyme [63]. LO involves T4 phage particles having lytic activity by possessing a lysozyme-like activity that happens at the base plate structure of the phage tail [64]. This is ineffective at a low multiplicity because cells develop an increasing resistance a few minutes after infection [65]. Lysis inhibition (LIN) can occur when the multiplicity of infection is between 5 to 10 and one or more phage particles has absorbed to the initially infected cells with each secondary absorption event causing a delay in lysis and cells accumulating intracellular phage for extended periods [66, 67]. The evolutionary advantages of both are evident as the rapid lysis of singly infected cells allow T4 to spread through a virgin population of cells and once the infection dominates LIN permits indefinite extensions of the latent period and extremely high progeny titres.

The main lysis genes in T4 bacteriophages are *e* and *t* that when mutated can prevent lysis without reducing the accumulation of intracellular phage particles [68].In the λ model T4 *e* is complementary to Lambdoid *R* and T4 *t* is complementary to Lambdoid *S* [69]. The *e* gene encodes the T4 lysozyme that has a globular structure and the lysozyme activity accumulates intracellularly from the onset of late gene expression, 8 min after infection [70]. The properties of *t* mutants are highly similar to those of λ *S* mutants and so it can be assumed that *t* functions in forming a hole in the inner membrane at the end of late-protein synthesis. LIN inhibition occurs only in cells in which "t-holes" have not yet been formed [66]. It is suspected that t-hole formation is directly controlled by the *r* complex.

In T7, the genes *17.5* and *3.5* in T7 play the same roles as the λ *S*/*R* and T4 *t*/*e* gene pairs [62] with gene *17.5* acting to form holes in the inner membrane through which the *3.5* encoded lysozyme can pass through [45] to rupture the cell wall.

1.10 Bacteriophage resistance mechanisms

Generally, bacterial cells protect themselves from phage infection with restriction-modification systems in which incoming phage DNA is degraded with restriction endonucleases and the host DNA is protected by methylation. Abortive infection systems become active if the restriction-modification system has failed. In *E. coli* the PrrC protein is a tRNA lys anticodon nuclease that causes cell death and prevents the phage from propagating and further infecting other cells [71]. It is also known that bacteria use phase variation to switch their restriction-modification systems on and off as it is thought that maintaining the methylation can affect global gene expression. A trade-off between growth and resistance is needed and when the restriction-modification system is turned off the bacteria may acquire virulence factors encoded by phage that undergo a lysogenic phase in their genome[71]. Another way that bacteria can protect themselves from infecting bacteriophages is with a CRISPR/Cas system, which requires prior exposure to a phage during which time a protospacer sequence from the phage is inserted between short palindromic repeats in the CRISPR system to become a spacer sequence that acts as a type of immune memory. Once the sequence is recognised again it expresses the Cas immune system that cleaves the invading DNA. However phages have also evolved ways around this by developing similar CRISPR/Cas systems themselves that can recognise spacers on inhibitory chromosomal islands produced by the bacterial host to restore phage replication[72].

There are many known bacteriophage resistance mechanisms that involve preventing phage adsorption or preventing phage DNA entry. In prevention of phage adsorption hosts can block their phage receptors, for example in *E. coli,* Phage T5 produces a lipoprotein to block its own receptors and prevent super-infection [73]. Hosts can also produce extracellular matrices to prevent phage adsorption by providing a physical barrier; Phage V10 of *E. coli* O157 can modify the O157 antigen to block adsorption because it has an O-acetyltransferase [74]. Sometimes phage receptors become blocked by competitive inhibitors; FhuA is an *E. coli* iron transporter that doubles up as a phage transporter for T1 and T5 but the phages can be outcompeted by microcin J25 [75]. In terms of preventing phage DNA entry, superinfection exclusion systems are used and coliphage T4 encodes these by *imm* and *sp* to inhibit injection of DNA and subsequent infection by other T-even-like phages [73].

1.11 Bacteriophage subversion of the host

Bacteriophages are also able to subvert the host through a number of mechanisms. T4 has IP(internal protein) genes that encode small, basic proteins that are encapsulated in the phage head and injected with the viral genome in early infection[76]. These IPs subvert host macromolecular biosynthesis and IPs can also counteract host defensive mechanisms against phage infection [77].

Phages are able to acquire host DNA modification in order to escape restriction by the R-M system and have their own anti-restriction systems. Phages can evade the R-M system by losing or modifying R-M recognition sites, co-injecting proteins that protect the phage DNA from restriction or inhibiting R-M enzymes [77]. For example, T7 is resistant to type III R-M enzymes because the recognition sites are in the same orientation and not in the head to head formation that is required for cleavage [78]. Also, T7 inhibits type I R-M enzymes with its Ocr (overcome classical restriction) protein that blocks the enzymes binding sites and is the first product to be expressed by T7 when it enters a bacterium [79].

1.12 Impact of understanding bacteriophage susceptibility and resistance in E. coli O157

Bacteriophages have successfully been used to kill bacteria in food and patients for decades with a successful period in the pre-molecular era of medicine that was quickly usurped with the invention of chemical antibiotics. Phage therapy is an important option in the fight against antimicrobial resistance as an alternative treatment for resistant bacterial infections [80]. A new appreciation for the importance of the human microbiota has also encouraged the renewed interest in phage therapy as it also has a potential role in modulating microbiota [81]. Phage therapy has been suggested to be less expensive and more specific than antibiotic treatment [82]. Recently, Phage therapy has been used in a randomized trial to treat bacterial diarrhoea [83] and to treat respiratory infections [84]. It also has applications in the food production industry with the recent use of phage for biocontrol of *Campylobacter jejuni* colonization in broilers [85]. To successfully implement bacteriophage therapy a better understanding of the strain's genetic arsenal against phage infection is needed. It is not only necessary to understand what makes a strain resistant but what also makes a strain susceptible to phages or groups of phages and understand what influences the change from one state to the other in order to anticipate it. It would be advantageous for the advancement of phage therapy to build up a set of gene markers to look for those that are defined in effecting bacteriophage resistance or susceptibility in order to be able to predict whether a sequenced strain will be resistant or susceptible to a given phage. Effectively this can be investigated using genome-wide association studies (GWAS) for bacteria in relation to phage susceptibility and resistance.

1.13 The public health impact of understanding phage typing

The world of microbiology is now in the genomics era.   It is likely that phenotypic typing such as phage typing will be phased out due to its low level of discrimination. In this transition period, it is valuable to make use of the extensive epidemiological data we have that is based on PT and move it into a genomic context by defining the genetic differences that account for PT.   Once that is understood, SNP-based typing for public health can be performed to cluster related strains while PT associated genes can still be detected in the sequences of the strains to link them to older outbreaks that may have been phage typed, i.e. backward compatibility. Moving forward the PT of a strain is likely to provide information about previous phage exposure and resistance, all improving our understanding of the molecular basis of bacterium-phage interactions. As antibiotic resistance provides a survival advantage, similarly increasing bacteriophage resistance may be an indicator of increased survival advantage in the environment. Those strains that are resistant to a wide array of different bacteriophages are less likely to be lysed. This is important for public health as it gives an indication of whether a strain is more or less likely to survive in different conditions and could potentially aid in the epidemiology of outbreaks.

1.14 Transposon-based genome-wide mutagenesis technologies

Genome-wide mutagenesis allows researchers to carry out screens for involved genes for a multitude of conditions with unprecedented depth in a wide range of bacterial species. Transposon-based genome-wide mutagenesis encompasses a wide range of variations on that technology, including Transposon insertion sequencing (TIS), transposon insertion site sequencing(Tn-Seq), insertion sequencing (INseq), high-throughput insertion tracking by deep sequencing (HITS) and transposon-directed insertion site sequencing (TraDIS) that all share the same fundamental methodology [86]. The basic methodology is summarised as follows: creation of a transposon library through transposon mutagenesis, pooling of the high-density library, growth of that library under desired growth conditions, transposon specific sequencing adaptors ligated to amplify transposon junctions for sequencing, high-throughput sequencing of insertions sites, map and count reads for each insertion site and look for loci with a change in insertions compared to the control. This change in insertions can either indicate an increased survival rate in that mutant (increased number of insertions compared to the control) or a decreased survival rate in that mutant (decreased number of insertions compared to the control). These techniques have been used successfully to identify essential genes in *Salmonella typhi* [87] and *Mycobacterium tuberculosis* [88]. TraDIS techniques have also been used to identify genes involved in bacteriophage infection [89].

1.15 Aims

- Sequence and annotate the 16 phages used to type STEC O157
- Sequence and annotate strains of STEC O157 submitted to the Gastrointestinal Bacterial Reference Unit from outbreaks and cases of sporadic gastrointestinal infection in the UK.
- Analyse the prophage-like elements of STEC O157
- Identify genetic elements associated with the PT phenotype
- Build on the current understanding of PT and its representation of clinical significance and environmental niche
- Identify STEC O157 genes associated with phage resistance or susceptibility by genome-wide mutagenesis

Chapter 2: Whole genome sequencing of the *Escherichia coli* O157:H7 typing phages

2.1 Background

The STEC O157 typing scheme utilises 16 typing phages. The aim of this chapter was to analyse the genome sequences of 16 (fourteen T4 and two T7) STEC O157 typing phages (TPs) to determine their relatedness in terms of gene content. Clustering based on differences in genome content aimed to reveal sub-clusters that could be used to simplify the typing scheme and identify genes that may account for differences in infectivity between related phages. Typing phages 5, 7 and 10 from the scheme have previously been sequenced [74, 90, 91]. The work presented in this first results chapter builds on these sequences by assembling the genomes of the majority of the remaining typing phages from short sequencing reads and then places the phage genomes into similarity groups and analyses how these relate to lytic activity and phage typing results.

A nestedness analysis is also described in this chapter which is a way of measuring the ranges of both host resistance and phage infectivity across a specialist to generalist gradient (Nestedness). I also performed a Modularity analysis which is the degree to which a network can be split into distinct modular groupings of phage and bacteria such that there are more infections within rather than between groups

2.2 Methods

2.2.1 Phage propagation and DNA extraction

The typing phages were obtained as a gift from the National Microbiology Laboratory, Winnipeg, MN, Canada to GBRU in the late 1980s. To propagate the phage, 0.1ml of the propagating strain (table 2.1, fig. 2.1) was inoculated into 2 x

20ml of single strength Difco nutrient broth. 0.1ml of test phage was added to one broth and the other kept as a control. The bottles were incubated and turbidity was monitored. When lysis was judged to be at its maximum compared to the control, a small amount of the phage solution was centrifuged at 2,200g for 20min. The supernatant was removed and spotted onto a flooded plate of propagating strain as a test; the plate was dried and incubated at 37°C overnight. The plates were examined for lysis (see fig. 2.2) and if positive the phage lysate was sterilized by filtration and stored at 4°C.

Table 2.1. Table showing phage types of the 16 propagating strains

Propagating strain for typing phage number	Phage type
PS 1	PT4
PS 2	PT4
PS 3	PT21/28
PS 4	PT14
PS 5	PT14
PS 6	PT14
PS 7	PT2
PS 8	PT14
PS 9	PT2
PS 10	PT32
PS 11	PT2
PS 12	PT14
PS 13	PT2
PS 14	PT14
PS 15	PT2
PS 16	PT4

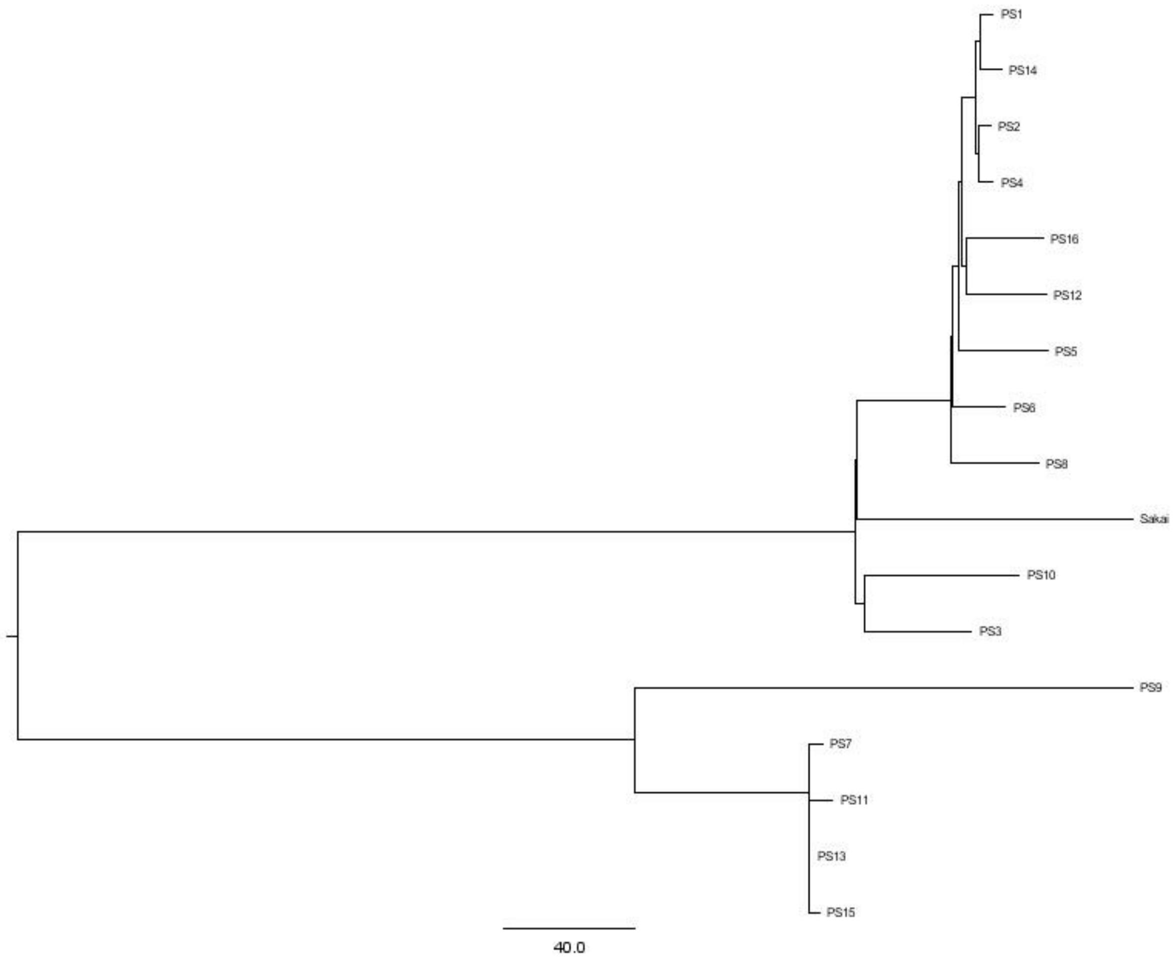


Figure 2.1. Phylogenetic tree of propagating strains for each typing phage using Sakai as a reference. SNPs were called using a mapping technique against Sakai and the tree drawn using MEGA 5.2. The labels stand for Propagating strain and then the corresponding phage number that they propagate. The data for this tree has been deposited in TreeBase with the submission ID 17186.

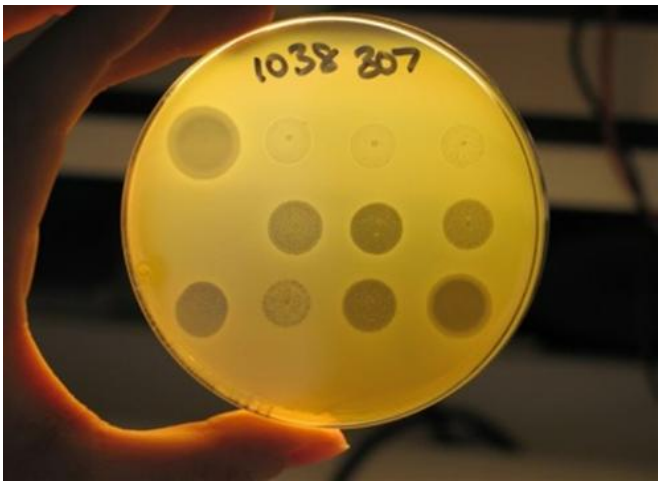


Figure 2.2 Example phage typing plate with lysis indicated by clear plaques on a lawn of bacteria, each plaque representing lysis for individual typing phages 1-16.

All phages were filtered before extraction took place. Eleven (phages 1, 3, 4, 5, 6, 7, 8, 9, 12, 13 and 14) of the 16 phages were extracted using the QIAamp UltraSens Virus kit (Qiagen, UK) following the manufacturer's instructions. This method failed to produce a high enough concentration of DNA for the remaining phages (2, 10, 11, 15 and 16) and these were extracted using a zinc chloride protocol [92]. Briefly, 20µl of a 2M zinc chloride solution was added to 1ml of sample and incubated for 5 min at 37°C. The sample was then centrifuged at 10,000rpm and the supernatant was removed. The pellet was suspended in 500µl of TES buffer (0.1M Tris-HCl, pH8; 0.1M EDTA and 0.3% SDS) and then incubated at 60°C for 15 min. Subsequently, 60µl of a 3M potassium acetate solution was added and the sample left on ice for 10 to 15 min. Following the formation of a white, dense precipitation the sample was centrifuged for 1 min at 12,000 rpm and the supernatant removed to a new tube. To this an equal volume of isopropanol was added, the solution vortexed and left on ice for 5 min. The solution was centrifuged and evaporated simultaneously using a Speedy-Vac machine and the pellet washed with 70% ethanol before being suspended in 20-100µl TE (10mM Tris-HCl, pH8; ImM EDTA). Samples were pooled by five extractions to give a higher yield of DNA. This method also failed to produce high enough concentration of DNA for sequencing TP 2 and 16 and so I was unable to obtain sequencing data for these two TPs.

2.2.2 Sequencing

The first set of phages (1, 3, 4, 5, 6, 7, 8, 9, 12, 13 and 14) was sequenced at The Genome Analysis Centre (TGAC) on an Illumina MiSeq. Illumina TruSeq DNA library construction was performed and sequencing of the libraries was pooled on one run using 150 bp paired-end reads, this generated greater than 1 Gbp of data for the run. The second set of phages (10, 11 and 15) was sequenced at the Animal Health and Veterinary Laboratories Agency on an Illumina GAII. The library construction was performed using a Nextera DNA sample preparation kit (Illumina) and then sequenced also producing 150bp paired reads.

2.2.3 Bioinformatic sequencing analysis

Reads for all phages apart from TP 15 were *de novo* assembled into whole genomes using Velvet optimizer with a range of k-mer values from 90-120 [93] and annotated using Prokka 1.5.2 and output as GenBank files [94]. The genomes were visualised in the multiple genome alignment tool Mauve with a progressive alignment to assess similarities and differences between them based on sequence content. The reads assembled into between 1 and 7 contigs for each phage. TP15 could not be assembled correctly because the propogation process had induced other temperate phages in the genome of the propagating strain and the DNA had been co-extracted. Subsampling to x150 coverage and assembling the reads using SPAdes with a low frequency k-mer elimination step [95] was used to overcome this issue and resolve 15 true TP15 contigs from the assemblies. The sequencing data has been made publicly available in the Short Read Archive under study alias PRJNA252693 and Genbank accession numbers for each phage can be found in the availability of supporting data section.

2.2.4 Euclidian tree

The protocol used to identify phage types (table 2.2) was converted into binary (presence/absence) format. In the original scheme there were 66 established phage types (PT) and 16 typing phages (TP). This set of data was analysed using a two-way cluster hierarchical agglomerative analysis in PC-ORD software version 6.08 (MJM software Design, Gleneden Beach, OR). The clustering was performed with Euclidian distance matrix and Ward linkage method.

The optimal number of groups of plots was first evaluated with multi-response permutation procedure, seeking the solution with fewest number of groups but the greatest gain in *A*-statistics [96].

Table 2.2 Reactions of the *E. coli* O157 type strains with the typing phages at a routine test dilution, this data was used to construct the Euclidian tree shown in Figure 2.3. Abbreviations are clear lysis (CL), semi-confluent lysis (SCL) and opaque lysis (OL). +++ = over 100 plaques, -/+ = 6-20 plaques, -/++ = 41-60 plaques, +++/- = 81-100 plaques, <OL = semi confluent opaque lysis, <<OL = less than semi confluent lysis,+++u = over 100 micro (μ) plaques

Phage Type	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	<CL	CL	SCL	SCL	SCL	-	<SCL	CL	-	-	CL	CL	-	<SCL	CL	SCL
2	CL	CL	<CL	<<SCL/-	-	<SCL	<SCL	CL	OL	-	CL	CL	CL	-	CL	SCL
3	+++	-	SCL	-	-	-	<SCL	-°	OL	OL	SCL	CL	-	-	SCL	SCL
4	CL	CL	SCL	<<SCL	<<SCL	-	<SCL	CL	OL	OL	CL	CL		<SCL	CL	<CL
5	-	-	SCL	SCL	-	SCL	SCL	-	OL	OL	<<OL	<OL	CL	-	-	-
6	-	-	SCL	SCL	-	+++	SCL	-	-	-	SCL	-	SCL	-	-	-
7	-	-	SCL	SCL	SCL	-	-	-	-	-	SCL	-	-	SCL	-	-
8	CL	CL	<CL	<<SCL	SCL	SCL	SCL	CL	-	-	CL	CL	CL	<SCL	CL	<SCL
9	SCL	SCL	SCL	CL	SCL	+++	SCL	SCL	-	-	SCL	CL	-	-	-	SCL
10	SCL	SCL	-	SCL	SCL	-	-	SCL	<<OL	<OL	+++	CL	-	CL	CL	CL
11	SCL	SCL	SCL	SCL	SCL	-	-	SCL	<<OL	-	SCL	SCL	-	SCL	SCL	SCL
12	SCL	SCL	SCL	CL	CL	-	SCL	SCL	-	-	+++	SCL	SCL	-	SCL	SCL
13	-	-	SCL	SCL	-	+++	+++	-°	<<OL	<<OL	<<OL	-	SCL	-	-	-
14	CL	CL	SCL	<<SCL	<SCL	<SCL	<<SCL	CL	OL	OL	CL	CL	CL	<<SCL	CL	<CL
15	SCL	+++	SCL	SCL	-	+++	SCL	SCL	-	-	SCL	SCL	SCL	-	SCL	SCL
16	-	+++	SCL	+++	-	+++	<SCL	-	<OL	<OL	-	+++	SCL	-	-	-
17	-	+++	+++	+++	-	-/+++	SCL	-	<OL	-	SCL	-	+++	-	-	-
18	-	-	+++	+++	-	-/++	+++	-	-	-	<<OL	-/++	<CL	-	-	-
19	-	+++	-	-	-	-	+++	-	-	-	<OL	+++	-	-	-	+++
20	-	-	+++	+++/-	-	++	-	-	-	-	-	-	<SCL	-	-	-
21	-	-	<<SCL	-	-	-	+++	-°	<OL	<OL	P -	-	-	-	-	-
21/28	-	-	CL	-	-	-	SCL	+++µ	<OL	<OL	+++µ	+++µ	-	-	-	-
22	-	-		-	-	-		+++	<OL	<OL		SCL	-	-	-	-
23	<CL	CL	SCL	-	-	-	SCL	CL	-	-	CL	<CL	-	-	CL	SCL
24	CL	CL	SCL	-	-	-	+++	CL	OL	-	CL	CL	-	-	CL	<<SCL

Phage Type	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
25	-	-	SCL	+++	-	<SCL	<SCL	-°	-	-	<<OL	<<OL	CL	-	-	-
26	<CL	SCL	+++	+++	+++	-	+++	CL	<OL	<OL	CL	CL	-	-	CL	SCL
27	+++	<<OL/-	SCL	+++/-	-	<<SCL	<<SCL	CL/-	+++	+++	CL	CL	CL	-	SCL	+++
28	-°	-	SCL	-°	-	-	<<SCL	-	<OL	<OL	<<OL°	<<OL°	-	-	-	-
29	<<OL°	-	+++	-	-	-	+++	+++	-	-	-	+++	-	-	+++	+++
30	++/-	<OL	+++	+++	+++	-	+++	-°	-	-	<<OL	<<OL	-	-	++/-	<OL
31	-	-	<SCL	-	-	<SCL	<SCL	-	-	-	P -	-	CL	-	-	-
32	-	-	<SCL	-	-	SCL	SCL	-°	OL	OL	P -	-	CL	-	-	-
33	-	-	<<SCL	-	-	-	<<SCL	-	-	-	P -	-	-	-	-	-
34	SCL	CL	+++	-	-	+++	+++	CL	<OL	<OL	CL	CL	<CL	-	CL	CL
35	CL	-°	<SCL	SCL	+++	-	+++	CL	-	-	CL	CL	-	-	<CL	-/++
36	<CL	-	<SCL	<<SCL	+++	-	+++	CL	<OL	<OL	CL	CL	-	-	CL	-°
37	CL	<<OL	SCL	+++	-/+	-	+++	CL	-	-	CL	CL	-	-	CL	<<OL
38	-	-	-	-	-	-	-	-	OL	OL	-	-	-	-	-	-
39	<<OL	-	+++	-	-	+++	+++	<OL	<OL	<OL	<<OL	<<OL	CL	-	<<OL	-
40	<OL	-	+++	-	-	-	+++	CL	OL	OL	CL	CL	-	-	SCL	-°
41	<CL	CL	<<SCL	<<SCL	+++	-	-	CL	OL	OL	CL	CL	-	+++	CL	<CL
42	<SCL	<<OL	<<SCL	-	-	-	+++	CL	<OL	<OL	SCL	SCL	-	-	<<SCL	
43	<<OL	-	SCL	-	-	<<SCL	+++	<<OL	-	-	<<OL	<<OL	CL	-	<<OL	
44	-°	<<OL	<<SCL	+++	+++	-	+++	P -	-	-	P -	-	-	-	P -	<SCL
45	<CL	<OL	+++	-	-/+	-	+++	CL	OL	OL	CL	CL	-	-/+	SCL	+++
46	-	-	<SCL	-	-	-	-	-	<OL	<OL	-	-	-	-	-	-
47	<CL	-	SCL	+++	+++	-	+++	CL	OL	OL	CL	CL	-	+++	CL	-
48	<CL	CL	SCL	<SCL	<<SCL	+++	-	CL	OL	OL	CL	CL	SCL	SCL	CL	SCL

Phage Type	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
49	<CL	-	SCL	<<SCL	<<SCL	<SCL	<CL	CL	OL	OL	CL	CL	CL	CL	CL	-
50	SCL	-	SCL	-	-	<<SCL	+++	CL	OL	-	CL	CL	CL	-	<OL	-
51	-	-	+++	-	-	+++	<<SCL	-	<OL	-	-/+	-	+++	-	-	-
52	SCL	-	+++	+	-	-	<<SCL	SCL	-	-	<CL	<<OL	-	++	<<SCL	-
53	<<OL	-	<<SCL	+++	+++	+++	<<SCL	<<OL	+++	<<OL	<<OL	<<OL	<<SCL	-	+++	+++
54	<CL	CL	+++	-	-	+++	++	<CL	-	-	<CL	SCL	<<SCL	-	<<SCL	<SCL
55	<CL	CL	<<SCL	<<SCL	+++	-	-	CL	-	-	CL	CL	SCL	SCL	CL	SCL
56	<CL	CL	+++	<<SCL	+++	-	-	CL	-	-	CL	CL	-	+++	CL	SCL
57	<<OL	<<OL	<<SCL	+++	+++	-	+++	<<OL	<OL	OL	-	-	-	-	+	SCL
58	-	<<OL	<<SCL	+	+	-	+++	+	-	-	-	-	-	-	-	SCL
59	-	+	<<SCL	+++	+++	+++	<<SCL	+	+++	<<OL	P -	P -	SCL	+++	-	<<SCL
60	-	+	+++	+++	+++	-	+++	-	-	-	P -	-	-	-	-	<SCL
61	SCL	<<OL	<<SCL	+++	+++	+++	-	CL	-	-	CL	CL	SCL	+++	SCL	+++
62	<CL	-	<<SCL	+++	++	-	<<SCL	-	+++	<OL	P -	-	-	+++	<CL	-
63	CL	SCL	-	-	-	-	SCL	CL	-	-	CL	CL	-	-	CL	SCL
64 (prov)	-	-	<<OL	+++	+++	-	+++	-	-	-	-	-	-	+++	-	<<SCL
65 (prov)	-	-	+++	<<SCL	+++	+++	<<SCL	-	<OL	<OL	P -	-	+++	-	-	+++
66 (prov)	-	-	<<SCL	-	-	+++	SCL	<<OL	<OL	<OL	P -	-	+++	-	-	-
67	CL	CL	<<SCL	-	-	-	-	CL	-	-	CL	CL	-	-	CL	<<SCL
68	CL	-	SCL	<<SCL	-	<<SCL	<<SCL	CL	OL	OL	CL	CL	<<SCL	-	<CL	-
69 (prov)	+++	-	+++	-	-	-	+++	-	-	-	-	-	-	-	-	+++
70 (prov)	CL	<<SCL	-	-	-	-	<<OL	CL	<<OL	<OL	CL	CL	-	-	CL	+++
71 (prov)	CL	-	+++	-	-	+++	<<SCL	CL	-	-	CL	CL	+++	-	<CL	+++
72 (prov)	-	-	+++	-	-	+++	-	-	<OL	<OL	-	-	++	-	-	-

Phage Type	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
73 (Not yet)																
74 (prov)	CL	<<SCL	-	-	-	-	-	CL	<OL	OL	CL	<CL	-	-	SCL	SCL
75 (prov)	<OL	-	+++	<OL	+++	+++	-	CL	-	-	CL	SCL	+++	<<OL	+++	-
76 (prov)	<<OL	-	<<SCL	+++	-	<<SCL	SCL	<<OL	<OL	OL	<<OL	<<OL	+++	-	-	-
77 (Not yet)																
78 (prov)	<<OL	-	<<SCL	-	-	-	<<SCL	<<OL	<OL	OL	<<OL	<<OL	-	-	-	-
79 (prov)	<<OL	-	-	-	-	-	<<SCL	<<OL	<OL	OL	<<OL	<<OL	-	-	<<OL	-
80 (prov)	<<OL	-	-	-	-	-	<<SCL	<<OL	<OL		<<OL	<<OL	-	-	-	-
81	<<OL	-	-	-	-	-	-	<<OL	<OL	<OL	<<OL	<<OL	-	-	<<OL	-
82	-	-	<<SCL	-	-	+++	<<SCL	CL	-	-	<CL	<<SCL	+++	-	-	-
88	+++	++	SCL	-	SCL	-	CL	SCL	-	-	+++	+++	-	-	+++	SCL
Untypable	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

2.2.5 Modularity and nestedness

Modularity of the network was calculated using the LPAb+ algorithm [97] which uses label propagation coupled with greedy multistep agglomeration to identify the communities (made of members of both types of nodes (bacteria and phage)) that maximise modularity in bipartite networks. As LPAb+ is stochastic, the best modularity score, Q_B , was returned from 1,000 trials each time the algorithm was used and was chosen. Code for performing the modularity analysis is supplied [98].

Nestedness statistics were calculated using FALCON [99]. The nestedness measures used were NODF [100], NTC [101, 102] and BR, the discrepancy score of Brualdi and Sanderson, 1999 [103]. NODF and NTC scores take values in the range [0,100], whilst BR is the absolute number of differences between the input and a maximally packed matrix. NODF has been recalculated here as NODF = 100-NODF, so that lower measure scores show greater nestedness with 0 representing perfect nestedness for each of the measures.

The significance of both modularity and the nestedness found in the phage-bacteria infection network using two null models based on properties of the network was examined. Null model one is a Bernoulli random null model where connections between phage j and bacteria i are made with probability $p_{ij} = F/M$, where F is the total number of edges in our network (number of infecting interactions) and M is the maximum number of potential interactions (number of TP's × number of PT's). Null model two is based on the information in the rows and columns in the network [104] ; where a connection between phage *j* and bacteria *i* is made with probability $p_{ij} = 0.5(d_j/r + k_i/c\,)$ where d_j is the number of infections caused by phage *j*, *r* is the number of PTs, k_i is the number of phage that can infect bacteria *i* and *c* is the number of TPs. One thousand null matrices against the network for each null model in the modularity analysis were tested, the adaptive ensemble of FALCON for nestedness analysis were used and the ensemble size used (N), p-values (probability of finding a more modular/nested network from the null model) and z-scores (effect size; the number of standard deviations the network was away from the mean average found in each null model) were reported.

2.2.6 BRIG plot

BRIG (Blast Ring Image Generator), a genome comparison tool [105], was used to compare similarities between the 12 T4 like TP by inputting all of the GenBank files for the assembled genomes and plotting blast hits against a FASTA file of all of the phages. The image was displayed as a series of concentric rings with the central ring being the FASTA reference; each outer ring displays hits (i.e. genomic regions that show a high percentage similarity to the central reference genome) for each phage. BRIG was also used to show the comparison of TP9 and TP10 (the two T7 like typing phages) against TP9 as a reference.

2.2.7 SeqFindR and Easyfig plots

SeqFindR, a bioinformatics tool developed by the Beatson Laboratory at the University of Queensland, was used to identify gene presence and absence in the phage genomes. Easyfig [106] was used to visualise the coding regions and colour the accessory genes in red for each phage group.

2.2.8 Tail fiber analysis

Tail fiber encoding genes were extracted from the GenBank files of the typing phages and the protein sequences aligned using MEGA 5.2. The alignment revealed how many changes in protein sequence there were within the groups.

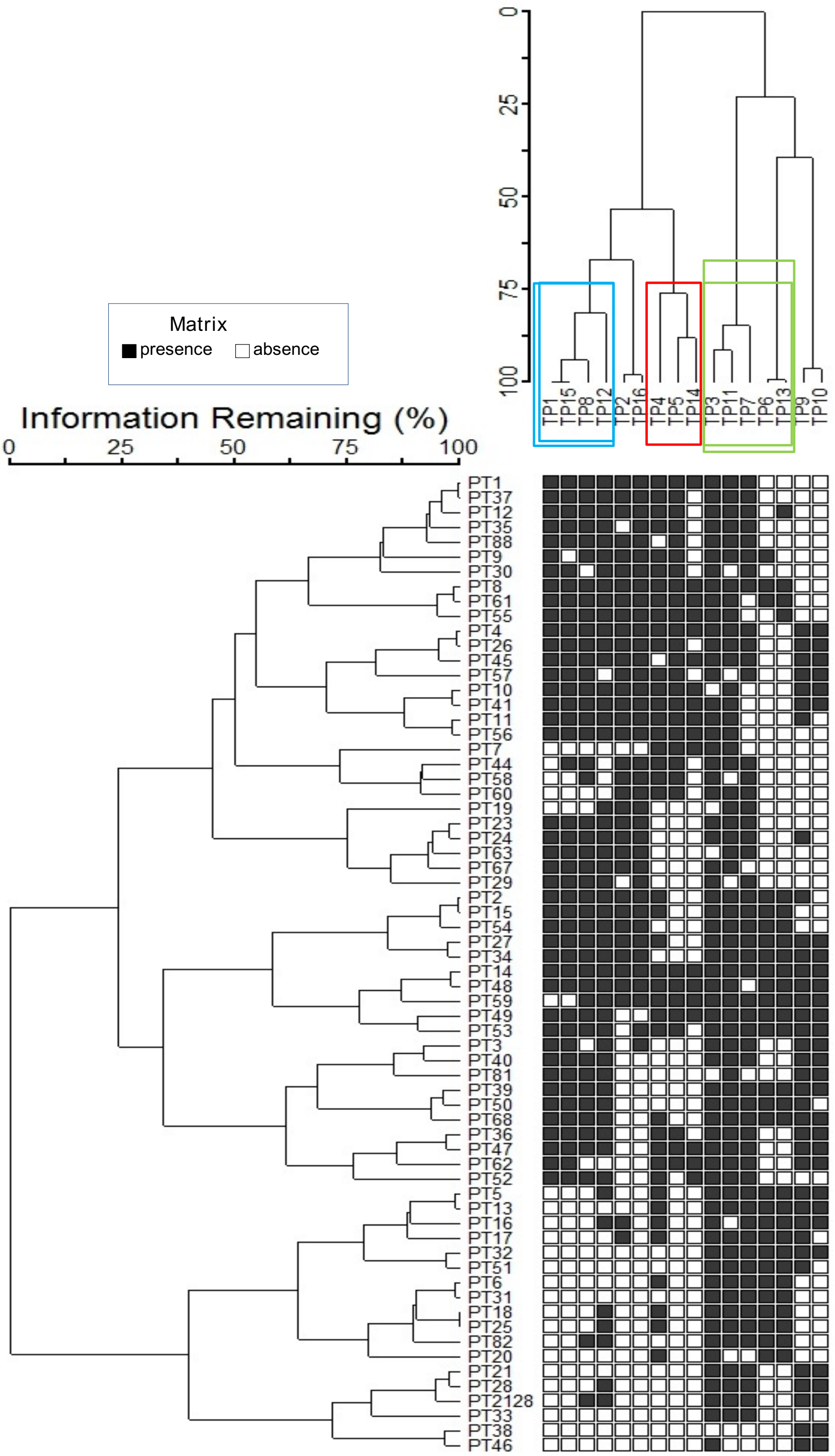
2.3 Results

2.3.1 Nestedness and modularity of the phage typing scheme

In the PT scheme there are 14 T4-like bacteriophages (TP1-8 and TP11-16) and two T7-like bacteriophages (TP9 and TP10). The reactivity of each of the typing phages with respect to the STEC O157 PT scheme was analysed. The two-way Euclidian cluster analysis combined the independent clustering of 66 STEC O157 bacterial PTs and the 16 TPs into a single diagram and highlighted the associations between groups of phage types and typing phages (Figure 2.3). The analysis showed that the STEC O157 phage typing scheme formed a weak (Q_b = 0.1575 (Table 2.3)) but significantly modular network where the TP groups were each specialised to infect a subset of PTs (Figure 2.4). There also exists a large number of the 'between module' interactions. Furthermore, the majority of PTs of STEC O157 are susceptible to at least one member of each group of TPs. These groups can be regarded as universally infective against STEC O157. Using statistical tests (defined in the descriptor of Table 2.3), the nestedness of the interaction network was found to be statistically significantly different from that found under randomly formed networks (Table 2.3, fig. 2.5). This indicates a correlation between phage infectivity range and the resistance range of the host. These phages have been chosen to infect STEC O157 and create a typing scheme with the simplest and minimum selection of phages so it makes sense that the system is nested.

Table 2.3 Summary statistics for nestedness and modularity analysis. Barber's modularity (Q_b) and three nestedness measures (NODF, NTC and BR) were calculated. Two null models were used to generate ensembles of networks (of size N) to evaluate the strength of the modularity and nestedness observed in the classified STEC O157:H7 phage-bacteria infection network. This is done by reporting the significance (as a *p*-value) and effect size (as a z-score) of the phage-bacteria infection network relative to the networks found in each null model ensemble. Note that, due to differences in how these measures are calculated, a positive z-score for modularity indicates that it is greater in the observed network than the mean average of the ensemble; whilst in the nestedness analysis a negative z-score indicates the observed network is more nested than the mean nestedness found within the null ensemble. The classified STEC O157:H7 phage-bacteria infection network was found to be both more nested and more modular than any of the networks generated by the tested null models.

		Modularity	Nestedness		
Measure		Q_B	NODF	NTC	BR
Measure score	x	0.1575	27.9199	30.2532	130
Null model 1	N	1000	1300	1300	1300
	p-value	<1/N	<1/N	<1/N	<1/N
	z-score	4.8602	-7.5382	-11.9831	-11.7632
Null model 2	N	1000	1000	1000	1000
	p-value	<1/N	<1/N	<1/N	<1/N
	z-score	5.7693	-4.6740	-6.7842	-7.1554



Matrix
■ presence
□ absence
Information Remaining (%)
0
25
50
75
100
TP1
TP15
TP8
TP12
TP2
TP16
TP4
TP5
TP14
TP3
TP11
TP7
TP6
TP13
TP9
TP10
PT1
PT37
PT12
PT35
PT88
PT9
PT30
PT8
PT61
PT55
PT4
PT26
PT45
PT57
PT10
PT41
PT11
PT56
PT7
PT44
PT58
PT60
PT19
PT23
PT24
PT63
PT67
PT29
PT2
PT15
PT54
PT27
PT34
PT14
PT48
PT59
PT49
PT53
PT3
PT40
PT81
PT39
PT50
PT68
PT36
PT47
PT62
PT52
PT5
PT13
PT16
PT17
PT32
PT51
PT6
PT31
PT18
PT25
PT82
PT20
PT21
PT28
PT2128
PT33
PT38
PT46

Figure 2.3 Two-way cluster analysis dendrogram of 66 phage types and 16 typing phages. The matrix of shaded squares represents the phage type × typing phage matrix, while the dendrograms show the clustering. The dendrograms are scaled by Wishart ‘s (1969) objective function, expressed as the percentage of information remaining at each level of grouping (McCune and Grace, 2002). Each square represents the presence (black) and absence (white) of a reaction with a given typing phage. The three phage type clusters and the 4 typing phage clusters are indicated at the node with numbers. The three T4 typing phage groups as indicated by their genomic relatedness and reactivity are represented with Group 1 in blue, Group 2 in green and Group 3 in red.

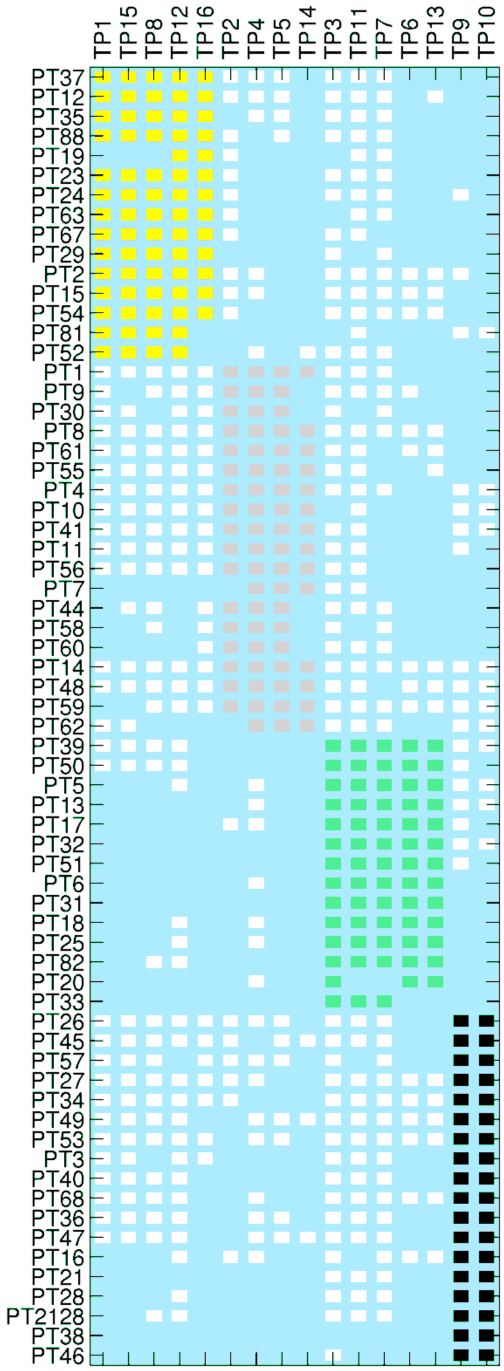


Figure 2.4 A visual representation of the modularity seen within the system with modules coloured. PT is represented on the y axis and TP is represented on the x axis and the matrix showing presence of a reaction with that phage as

a white or coloured block. The 4 observed modules are coloured as yellow, pink, green and black.

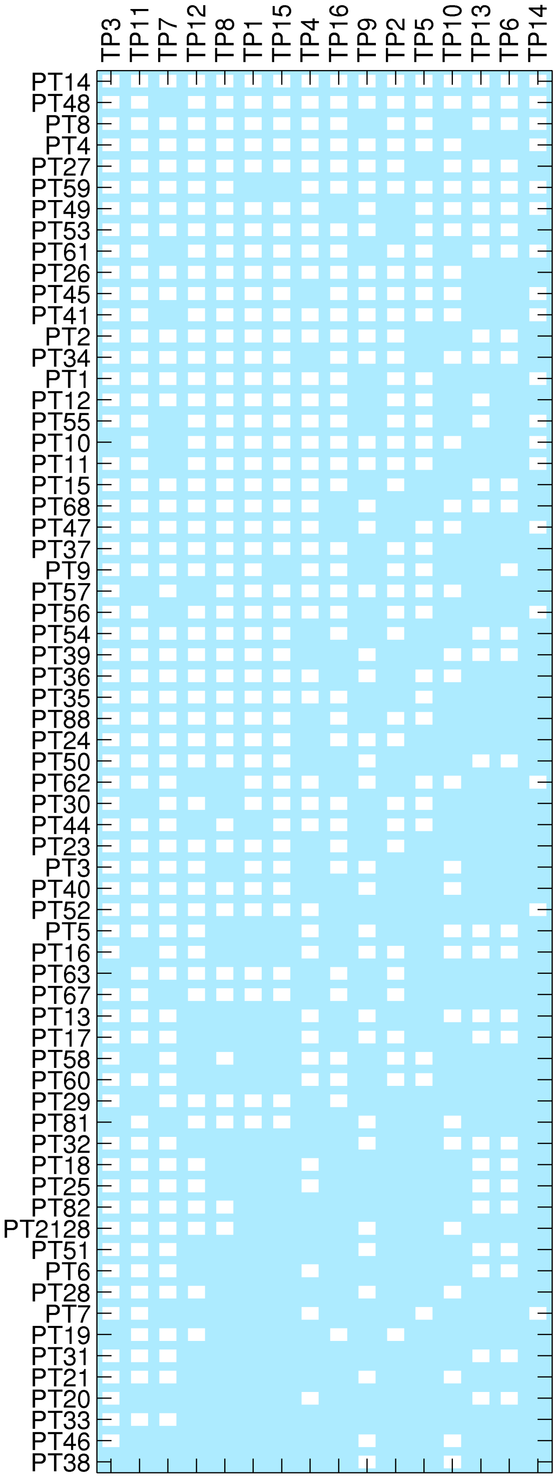


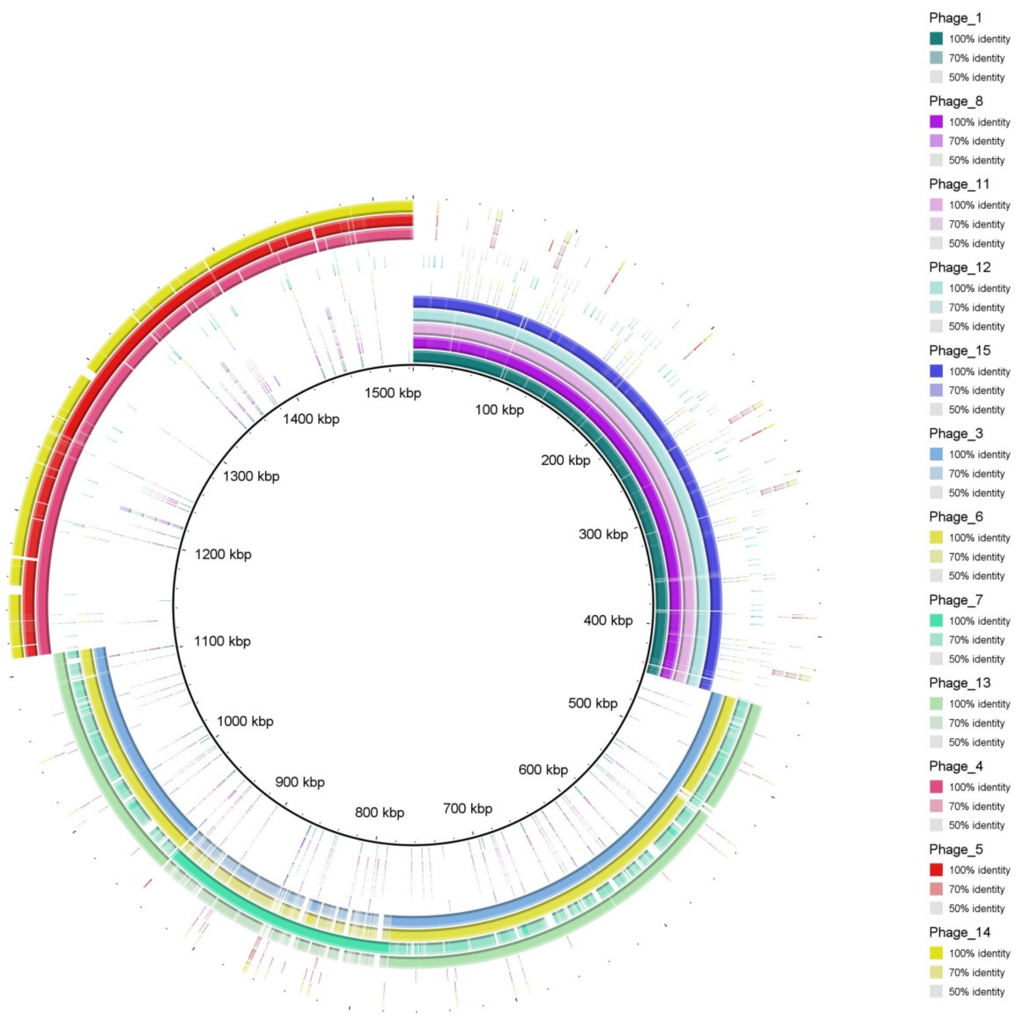
Figure 2.5 A visual representation of the degree of nestedness found within the classified *Escherichia coli* O157:H7 phage-bacteria infection network. Each square represents an association between the corresponding phage and bacterial strains. TP are ranked from highest to lowest phage infectivity range, whilst bacterial PT are ordered from lowest to highest host resistance range.

2.3.2 Fourteen of the 16 typing phages were successfully sequenced

Fourteen of the 16 phages in the typing scheme were sequenced and successfully assembled. Despite several attempts, sequencing of TPs 2 and 16 failed due to insufficient quantities of DNA extracted from the phage preparations.
Details of public accessions for both the raw reads and assembled genbank files can be found in the appendix.

2.3.3 Genomic similarity of the T4-like typing phages

The BRIG plot showed that the 12 sequenced T4-like bacteriophages formed three distinct groups of similar genomic sequences (fig. 2.6). Group 1 included typing phages 1, 8, 11, 12 and 15; Group 2 comprised typing phages 3, 6, 7 and 13 and typing phages 4, 5 and 14 were in Group 3. Although the sequencing for TP2 and TP16 failed, the modularity analysis indicates that TP16 belonged to Group 1 and TP2 belonged to Group 3 (fig. 2.6). The genomes of the TPs varied significantly in size between the three groups: the members of Group 1 were 93,000-95,000 bp, Group 2 members were 165,000-175,000 bp and those in Group 3 were 135,000-140,000 bp.



.

Figure 2.6 A genomic representative diagram drawn with BRIG of 14 T4-like phage similarities, the coloured regions indicate high pairwise genomic sequence similarity according to blastn. Legend indicates which colours correspond with which phages and the shade of that colour indicates what level of similarity is observed. Central ring is multi-fasta of all T4-like phage genomes and each consecutive ring represents the similarity with a single phage. The multi-fasta and rings are in the same phage order. The reference (central ring) is a concatenated

sequence that permits visualisation of conservation of sequences for each of the three phage groups.

2.3.4 Group 1 variation

The Group 1 phages (TP1, 8, 11, 12 and 15) were approximately 90,000 bp in length. These five phages were highly similar in genetic sequence content. The location, annotation and presence of accessory genes within Group 1 are shown in figure 2.7 and table 2.4. Figure 2.7 shows that there were 6 genes found in TP1 but absent in TP8, 11, 12 and 15 (five were annotated as hypothetical proteins and one tRNA). There were also five genes present in TP8, 11, 12 and 15 but not in TP1 (three were annotated as hypothetical proteins, one as AP2 domain protein and one was a tRNA gene) (fig. 2.7, table 2.4). TP8 was missing a region annotated as a putative prophage that was present in the other members of the group. With the exception of TP11, the Group 1 TPs are most closely related to each other by the two-way Euclidian cluster analysis demonstrating the link between gene content and phage typing profile.

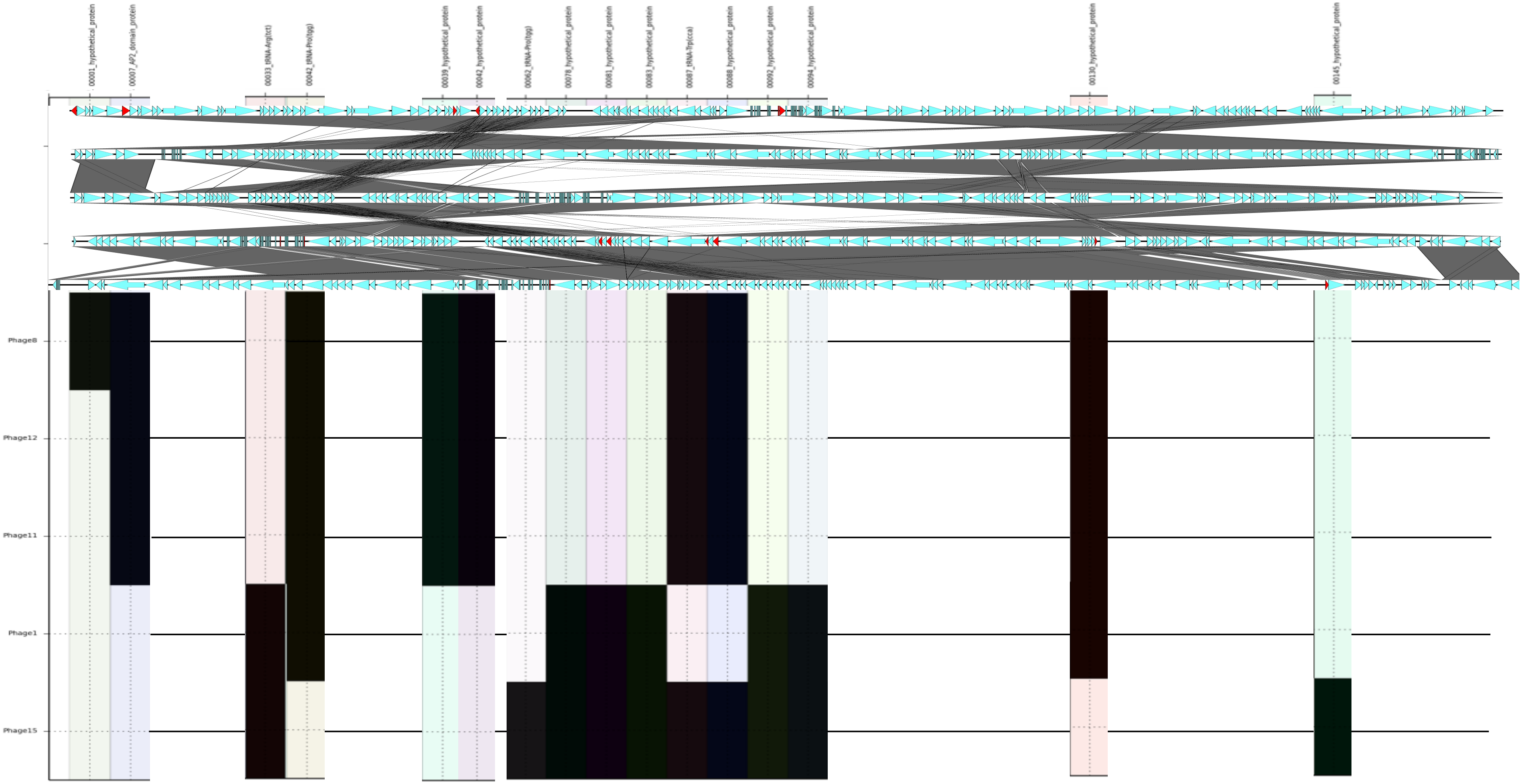

00001_hypothetical_protein
00007_AP2_domain_protein
00033_tRNA-Arg(tct)
00042_tRNA-Pro(tgg)
00039_hypothetical_protein
00042_hypothetical_protein
00062_tRNA-Pro(tgg)
00078_hypothetical_protein
00081_hypothetical_protein
00083_hypothetical_protein
00087_tRNA-Trp(cca)
00088_hypothetical_protein
00092_hypothetical_protein
00094_hypothetical_protein
00130_hypothetical_protein
00145_hypothetical_protein
Phage8
Phage12
Phage11
Phage1
Phage15

Figure 2.7 SeqFindR and Easyfig image combined representing the accessory gene content of group 1. Genomes of each phage in group 1 are represented by the Easyfig image showing linear visualisation of the genome and coding regions represented by arrows, accessory genes are coloured red. The order of phage genomes in the linear visualisation and the accessory content blocking is 8, 12, 11, 1 and 15 and was chosen based on similarity clustering in SeqFindR. Hits for the accessory genes in each genome are represented in labelled columns in the SeqFindR image underneath each accessory gene.

Table 2.4 Table detailing the accessory variation of Group 1 as depicted in Figure 2.7. Presence of accessory gene in phage genome depicted by 'X' and absence by a blank box.

Group 1	Phage 1	Phage 8	Phage 11	Phage 12	Phage 15
PROKKA1_00033_tRNA-Arg(tct)	X				X
PROKKA1_00042_tRNA-Pro(tgg)	X	X	X	X	
PROKKA1_00078_hypothetical_protein	X				X
PROKKA1_00081_hypothetical_protein	X				X
PROKKA1_00083_hypothetical_protein	X				X
PROKKA1_00092_hypothetical_protein	X				X
PROKKA1_00094_hypothetical_protein	X				X
PROKKA1_00130_hypothetical_protein	X	X	X	X	
PROKKA8_00001_hypothetical_protein		X			
PROKKA8_00007_AP2_domain_protein		X	X	X	
PROKKA8_00039_hypothetical_protein		X	X	X	
PROKKA8_00042_hypothetical_protein		X	X	X	
PROKKA8_00087_tRNA-Trp(cca)		X	X	X	X
PROKKA8_00088_hypothetical_protein		X	X	X	X
PROKKA15_00062_tRNA-Pro(tgg)					X
PROKKA15_00145_hypothetical_protein					X

2.3.5 Group 2 variation

The TPs in Group 2 (TP 3, 6, 7, and 13) were between 160-170,000 bp in length. The genomes were almost twice the size of the phages in Group 1 and exhibited less similarity in terms of gene content. The accessory genes found in Group 2 were mostly annotated as encoding hypothetical proteins (fig. 2.8, appendix table 7.1). The two-way Euclidian cluster analysis highlighted a close relationship between TP6 and TP13 and this corresponded with the level of sequence similarity of these two TPs illustrated in figure 2.8.

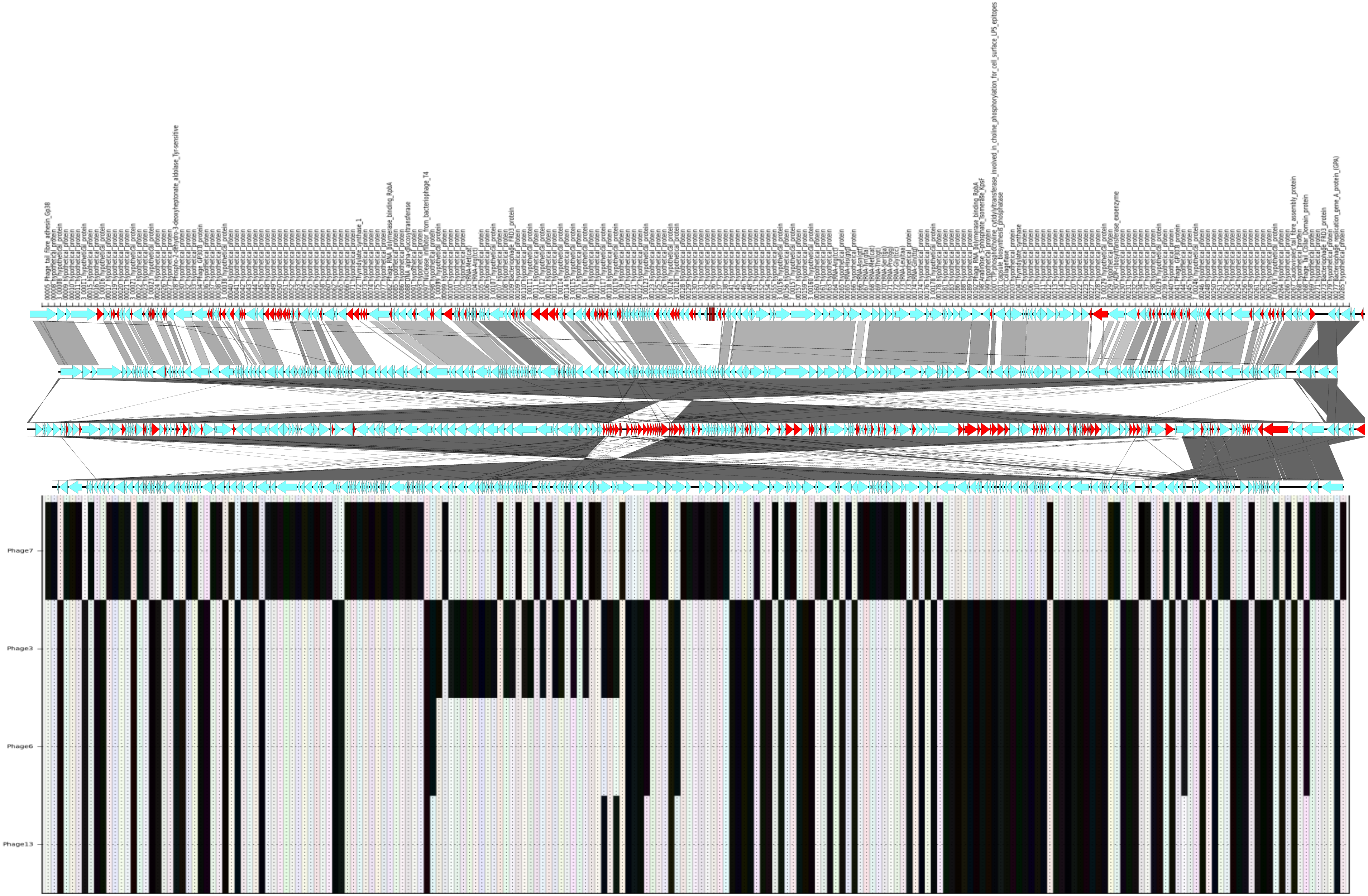

Phage7
Phage3
Phage6
Phage13

Figure 2.8 SeqFindR and Easyfig image combined representing the accessory gene content of group 2. Genomes of each phage in group 2 are represented by the Easyfig image showing linear visualisation of the genome and coding regions represented by arrows, accessory genes are coloured red. The order of phage genomes in the linear visualisation and the accessory content blocking is 7, 3, 6 and 13 and was chosen based on similarity clustering in SeqFindR. Hits for the accessory genes in each genome are represented in labelled columns in the SeqFindR image underneath each accessory gene.

2.3.6 Group 3 variation

TPs 4, 5 and 14 were designated Group 3 and were 130-140,000 bp in length. Figure 2.9 shows the location, annotation and presence of accessory genes within Group 3. Figure 2.9 demonstrates that there were 29 gene differences within the group and the majority (19) were annotated as hypothetical proteins. In addition, three genes encoded putative endonucleases and there were three genes designated tRNAs (figure 2.9, table 2.5). The TPs in Group 3 were most closely related to each other by the two-way Euclidian cluster analysis (fig. 2.3).

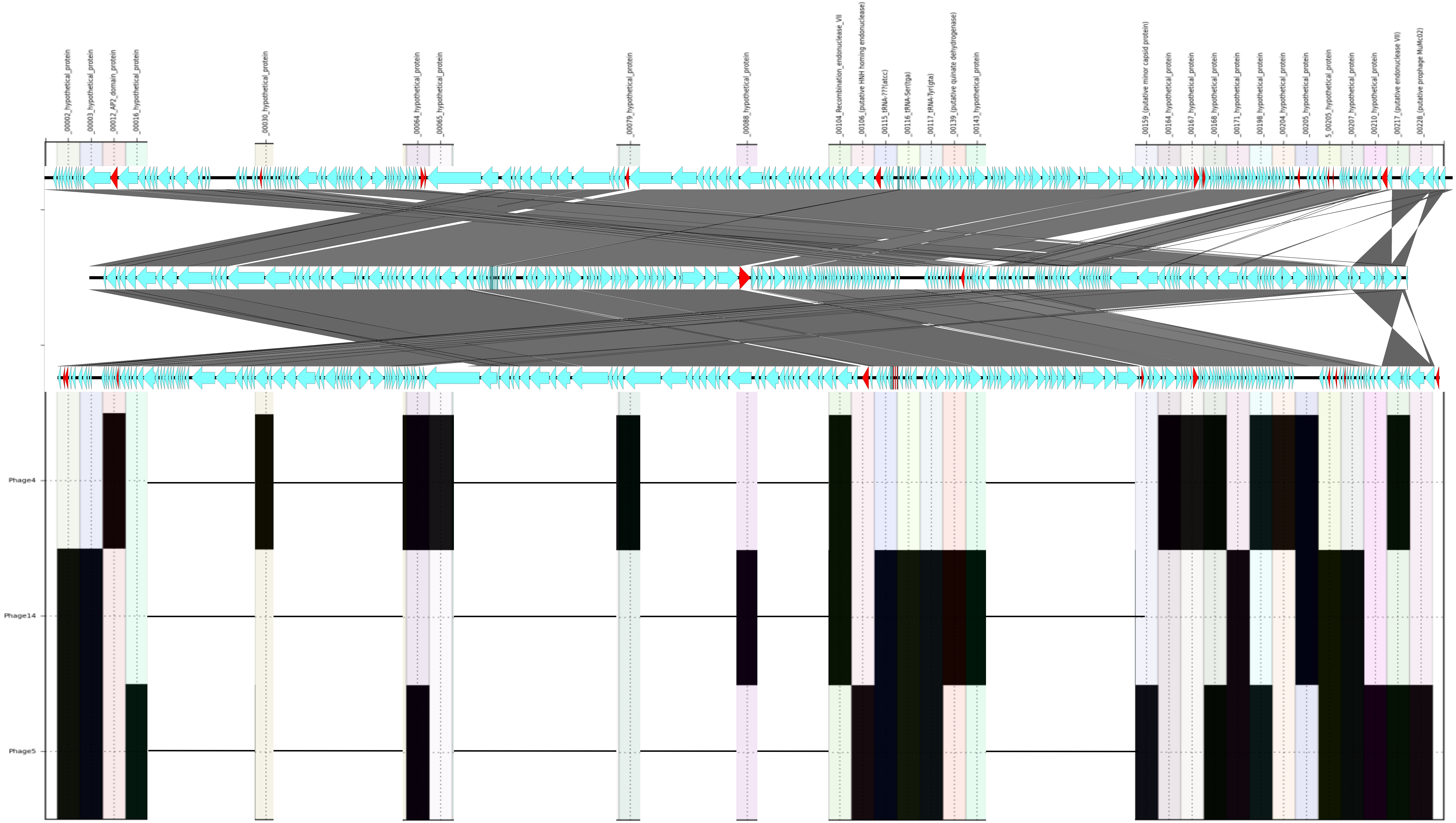

_00002_hypothetical_protein
_00003_hypothetical_protein
_00012_AP2_domain_protein
_00016_hypothetical_protein
_00030_hypothetical_protein
_00064_hypothetical_protein
_00065_hypothetical_protein
_00079_hypothetical_protein
_00088_hypothetical_protein
_00104_Recombination_endonuclease_VII
_00106_(putative HNH homing endonuclease)
_00115_tRNA-???(atcc)
_00116_tRNA-Ser(tga)
_00117_tRNA-Tyr(gta)
_00139_(putative quinate dehydrogenase)
_00143_hypothetical_protein
_00159_(putative minor capsid protein)
_00164_hypothetical_protein
_00167_hypothetical_protein
_00168_hypothetical_protein
_00171_hypothetical_protein
_00198_hypothetical_protein
_00204_hypothetical_protein
_00205_hypothetical_protein
5_00205_hypothetical_protein
_00207_hypothetical_protein
_00210_hypothetical_protein
_00217_(putative endonuclease VII)
_00228_(putative prophage MuMc02)
Phage4
Phage14
Phage5

Figure 2.9 SeqFindR and Easyfig image combined representing the accessory gene content of group 3. Genomes of each phage in group 3 are represented by the Easyfig image showing linear visualisation of the genome and coding regions represented by arrows, accessory genes are coloured red. The order of phage genomes in the linear visualisation and the accessory content blocking is 4, 14 and 5 and was chosen based on similarity clustering in SeqFindR .Hits for the accessory genes in each genome are represented in labelled columns in the SeqFindR image underneath each accessory gene.

Table 2.5 Table detailing the accessory variation of Group 3 as depicted in Figure 2.9. Presence of accessory gene in phage genome depicted by 'X' and absence by a blank box.

Group 3	Phage 4	Phage 5	Phage 14
PROKKA5_00002_hypothetical_protein		X	X
PROKKA5_00003_hypothetical_protein		X	X
PROKKA4_00012_AP2_domain_protein	X		
PROKKA5_00016_hypothetical_protein		X	
PROKKA4_00030_hypothetical_protein	X		
PROKKA4_00064_hypothetical_protein	X	X	
PROKKA4_00065_hypothetical_protein	X		
PROKKA4_00079_hypothetical_protein	X		
PROKKA14_00088_hypothetical_protein			X
PROKKA4_00104_Recombination_endonuclease_VII	X		X
PROKKA5_00106_(putative HNH homing endonuclease)		X	
PROKKA5_00115_tRNA-???(atcc)		X	X
PROKKA5_00116_tRNA-Ser(tga)		X	X
PROKKA5_00117_tRNA-Tyr(gta)		X	X
PROKKA14_00139_(putative quinate dehydrogenase)			X
PROKKA14_00143_hypothetical_protein			X
PROKKA5_00159_(putative minor capsid protein)		X	
PROKKA4_00164_hypothetical_protein	X		
PROKKA4_00167_hypothetical_protein	X		
PROKKA4_00168_hypothetical_protein	X	X	
PROKKA5_00171_hypothetical_protein		X	X
PROKKA4_00198_hypothetical_protein	X	X	

PROKKA4_00204_hypothetical_protein	X		
PROKKA4_00205_hypothetical_protein	X		X
PROKKA5_00205_hypothetical_protein	X	X	
PROKKA5_00207_hypothetical_protein	X	X	
PROKKA5_00210_hypothetical_protein		X	
PROKKA4_00217_(putative endonuclease VII)	X	X	
PROKKA5_00228_(putative prophage MuMc02)		X	

2.3.7 T7 variation

TP9 and 10, the two Podoviridae or T7-like phages in the typing scheme were successfully sequenced, assembled and annotated and revealed 40-45000 bp genomes consistent with the published sequences of T7 bacteriophages (fig. 2.10). Phages 9 and 10 differed by only three genes (annotated as encoded hypothetical proteins) that were found in Phage 9 but not in phage 10. The two-way Euclidian cluster analysis confirmed the close relationship between TP9 and TP 10 in terms of PT profile. It also showed that there were six STEC O157 phage types (PTs 2, 11, 17, 24, 50, and 51) that react with TP9 but not TP10 and none of the PTs react with TP10 but not TP9 (fig. 2.3). These three hypothetical proteins could be the key to the differences in the reactivity profiles of TP9 and 10.

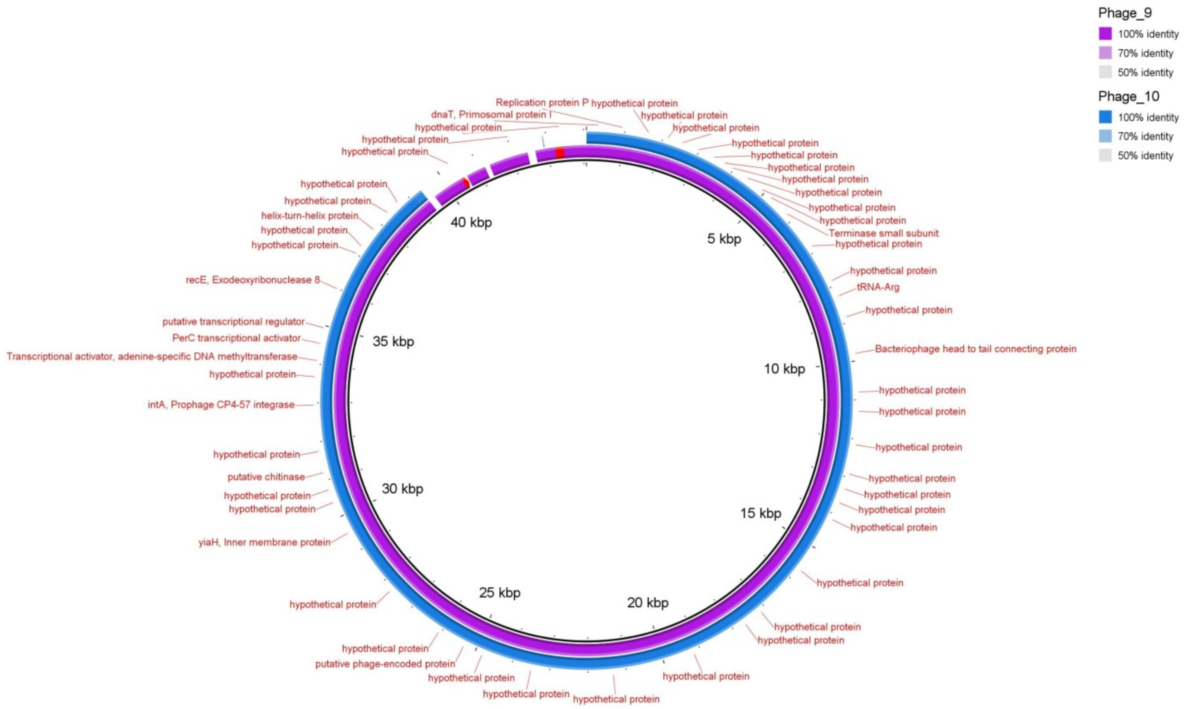


Figure 2.10 A genomic representative diagram drawn of T7 bacteriophage with BRIG, the coloured regions indicate high pairwise genomic sequence similarity according to blastn. The legend indicates which colours correspond with which phages. The central ring is a GenBank file of Phage 9 as a reference and annotations of genes in red. The first ring is representative of Phage 9 and the second ring is representative of Phage 10 and the shade indicates the level of genomic similarity observed.

2.3.8 Tail-fiber variation

Bacteriophages are known to have their specificity determined by their tail fibers which interact with receptors that are outer membrane proteins [107] on the host. The T4-like bacteriophages are long tailed (fig. 2.11) and the T7-like bacteriophages are short-tailed (fig. 2.12). Each group of the T4 phages had one tail fiber gene and one tail fiber assembly gene or tape measure domain annotated gene. The T7 phage groups had head-to-tail connecting proteins as they are short-tailed phages. Tail-fiber encoding genes were analysed within each group and it was found that there were nucleotide sequence polymorphisms that would result in changes to the amino acid sequence for certain members of each group. Within the group 1 typing phages, phages 1 and 15 had three changes in predicted amino acid sequence in their tail fibers, two of which were shared and one each unique to each phage. Within the group 2 typing phages, phage 7 has 47 predicted changes in its amino acid sequence and three amino acid deletions. Within the group 3 typing phages, the same single position in all three members of the group has a different predicted amino acid present and additionally there was another single nucleotide polymorphism in typing phage 14 that results in an amino acid change. The T7-like phages had identical tail fiber genes.

There was no genetic similarity between tail fiber genes found in different groups as they were non-homologous sequences. It was expected that the T4 phages that infect *E. coli* O157 would have similar tail fibers between groups as a function of convergent evolution but the fact that they do not is further evidence of the modularity of the PT scheme and that each group is particularly specialised and not closely related to the other groups. This also indicates that either there are multiple receptors for the different tail fibers on O157 or that multiple tail fibers recognise the dominant receptors on the STEC O157 clone.

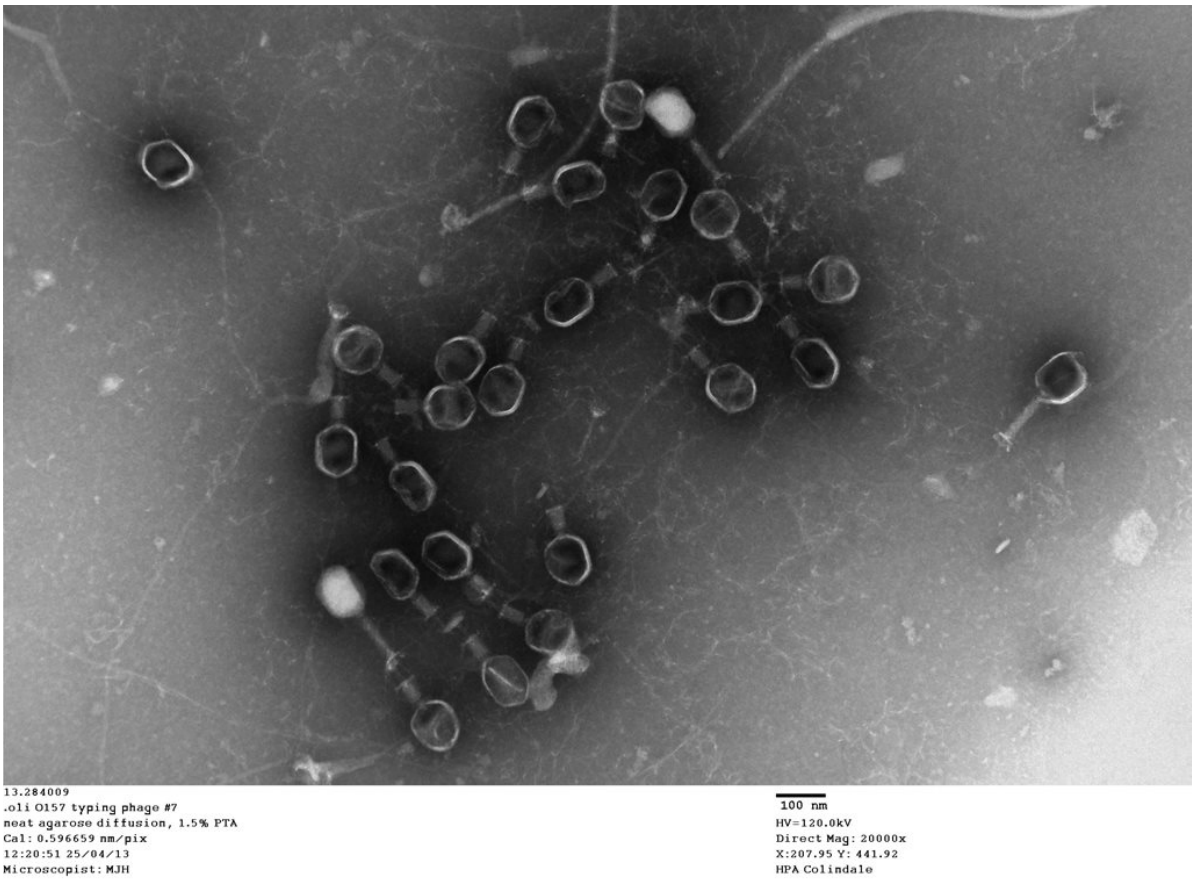


Figure 2.11 Electron Microscopy was performed image of typing phage 7, a representation of T4-like long-tailed phage morphology. The samples were applied to formvar-carbon coated EM grids by the agarose diffusion method and negatively stained with 1.5% PTA. EM grids were viewed at 120kV in a JEM1400 transmission electron microscope (JEOL UK) and digital images acquired using an AMT XR60 camera (Deben UK).

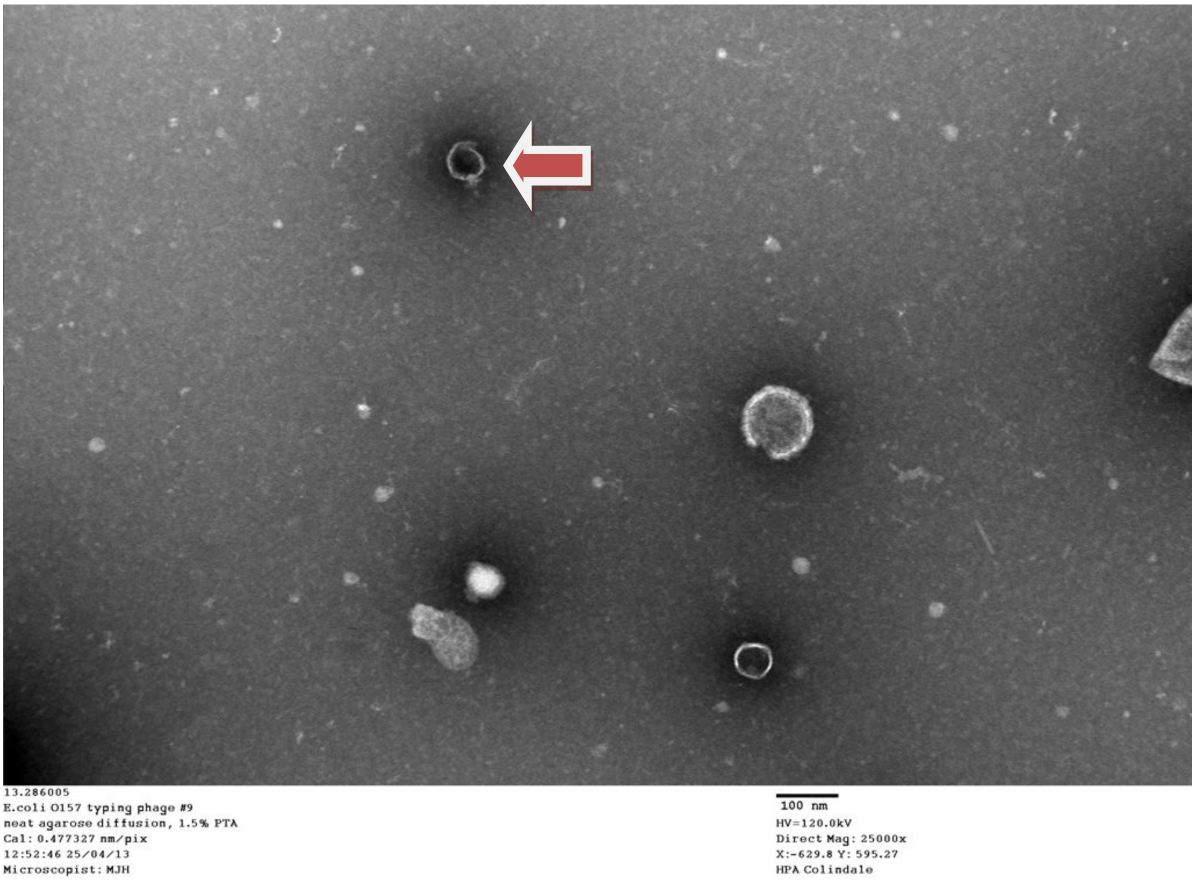


Figure 2.12 Electron Microscopy image of typing phage 9, a representation of T7-like short-tailed phage morphology. This image shows two phages and one is indicated by a red arrow.

2.4 Discussion

Phage-host interactions are key to understanding the virulence and success of STEC O157 but little is known about the TPs used in the O157 typing scheme. Sequencing these phages has enabled the grouping of the T4-like Myoviridae and the two Podoviridae or T7-like phages members of the TP scheme into four groups based on their sequence similarity. The two-way Euclidian cluster analysis demonstrated that similar phage groups react with STEC O157 in a similar way with closely related reaction profiles. The sequencing data also highlighted that a small number of gene differences may be responsible for the subtle variation in reaction profiles within the groups.

The large proportion of genes annotated as encoding hypothetical proteins hindered the investigations into the mechanisms of host-phage interactions. Attempts were made to annotate these genes further using protein BLAST and HMMER but only uncharacterised proteins were hit. However, the determination of which genes vary within each group will enable researchers to focus on the genes that may play a key role in the interactions between specific TPs and strains belonging to specific PTs. For example, in Group 1, there were five genes that were found only in TP8, 11 and 12 and three PTs (PT21/28, 59 and 82) that only react with these TPs. The proteins encoded by these five genes may play a key role in the host-phage interactions between TP8, 11 and 12 and strains of STEC O157 belonging to PT21/28, 59 and 82. PT21/28 is one of the most common PTs in the UK and is significantly associated with HUS [108]. Further details of unique host-phage interactions are described in table 2.6 and the genes referred to can be found in figures 2.7, 2.8 and 2.9.

Table 2.6 Table representing unique reactions that only occur within a subset of groups 1, 2 and 3 with specific PTs and number of genes found only in that subset. The number of genes was determined looking at the accessory gene regions within each group and identify those found in a subset of the phages.

Group	**Members**	**PTs that react with subsets of the group**	**No. of gene differences for that subset**
1	Single member of group 1	TP8 reacts with PT58 TP11 reacts with PT6,7, 13,17, 21,31,32,33,51,60 TP12 reacts with PT16 TP15 reacts with no PTs	1 0 0 2
	Two members of group 1	TP11 and 12 react with PT5,18, 19,25,28 TP1 and 15 react with no PTs	0 6
	Three members of group 1	TP8, 11 and 12 react with PT21/28,59,82 TP1, 15 and 11 react with PT62 TP11, 15 and 8 react with PT44 TP1, 15 and 12 react with PT30 TP1, 15 and 8 react with PT57	5 0 0 0 0
	Four members of group 1	TP1, 8, 11 and 12 react with PT9 TP1, 8, 15 and 12 react with PT29 TP1, 11, 12 and 15 react with PT3 TP8, 11, 12 and 15 react with no PTs	2 0 0 2
	All	**1,2,4,8,10,11,12,14,15,23,24,26,27,34,35,36,37,39,40,41,45,47,48,49,50,52,53,54,55,56,61,63,67,68,81,88**	

	None	**20,38,46**	
2	Single member of group 2	TP3 reacts with PT3,7,11,40,41,56,67 TP7 reacts with PT19,63	19 107
	Two members of group 2	TP3 and 6 react with no PTs TP3 and 7 react with PT1,4,21,21/28,23,24,26,28,29,30,33,35,36,37,44,45,47,52,57,58,60,62,88 PT3 and 13 react with PT55	5 0 2
	Three members of group 2	TP3, 6 and 7 react with PT9 TP3, 7 and 13 react with PT12 TP6, 7 and 13 react with PT46 TP3, 6 and 14 react with PT20,48, 61	0 0 0 80
	All	**2,5,6,8,13,14,15,16,17,18,25,27,31,32,34,39,49,50,51,53,54,59,68,82**	
	None	**10,38,81**	
3	Single member of group 3	TP4 reacts with PT2,5,6,13,15,16,17,18, 20,25,27,68 TP5 reacts with PT88 TP14 reacts with no PTs	7 5 3
	Two members of group 3	TP4 and 5 react with PT12,26,35,36,37,44,53,57,58,60 TP4 and 14 react with PT52 TP5 and 14 react with PT45	4 2 8

	All	**1,4,7,8,10,11,14,41,47,48,49,55,56,59,61,62**	
	None	**3,19,21,21/28,23,24,28,29,31,32,33,34,38,39,40,46,50, 51,54,63,67,81,82**	

Analysis of tail fibers genes showed that typing phages 1, 15, 7 and each individual member of Group 3 had different protein sequences encoded to the other members of their group. The changes that were found could partially account for infectivity differences [109]. These could explain a few of the differences in host specificity seen within those groups, although this will not apply to the T7-like TPs that have had identical predicted tail fiber proteins.

Certain TPs had almost identical genomes but different host susceptibility profiles, for example, TP11 belonged to the Group 1 phages but had a similar host susceptibility profile to the Group 2 phages. Each phage in the typing scheme has its own propagating strain (fig. 2.2 and table 2.1) so it is also possible that host-induced modification occurs [110]. For example, the propagating strains for the closely related TPs TP9 and TP10 are STEC O157 PT2 and PT32, respectively. Modifications may be a result of methylation or other phenotypic changes that are not evident in the genome but may affect the host range of the virus.

Phenotypic differences in susceptibility patterns in genetically similar phages could be explained by the transcription order of genetic loci in the phage genome. It has been suggested that gene synteny constrains adaptation and is important for fitness and therefore infectivity of bacteriophages [111]. The order of transcription may be important in overcoming the host response to infection. The phages that transcribe their genetic loci in a different order may be killed and degraded by the host response, for example, TP 8, 11 and 12 are almost identical but have a different gene order and this may be key to their different infection profiles.

This analysis has shown that the significantly modular network exhibited by the STEC O157 PT scheme was linked to the genetic similarity groups described above showing that these groups are specialised to infect a subset of PTs. However, the typing scheme as a whole is also significantly nested; more generalised phages minimise the number of phages needed in the scheme. Both of these network structures have also been found in other phage-bacteria infection networks [112, 113]. The most common PTs in the UK: 2, 8, 21/28 and 32 are all found in different modules, meaning there is an abundant PT in each module. When looking at these PTs with nestedness, PT 8 and 2 both have a phage susceptibility range of 14 and 13 respectively so are relatively generalised but PT 21/28 and 32 both have a host range of 7, and lie more towards the specialised end of the spectrum. It is interesting that the most common PTs detected in the UK seem to appear at two levels of host range – perhaps suggestive of a trade-off between host range and phage productivity. It would be interesting to see, in conditions where the phages are allowed to evolve with their hosts, if a more modular network arises with further specialisation of the phages to maintain a kill-the-winner dynamic and less broad range infectivity [114]. This is an artificial system that has been observed and it is likely that different network would arise in nature's ecological systems.

PT has been used for epidemiological and surveillance studies by a number of groups [43, 115] for different organisms. PT association with increased strain virulence is of high interest to public health workers dealing with STEC O157 and the replacement of phage-typing with whole genome sequencing should incorporate our knowledge of phage type and associated virulence. For this reason it is valuable to find the molecular markers associated with high frequency and highly pathogenic phage types; elucidating the determinants underpinning differences in phage typing should contribute to this.

Phage-mediated therapies will continue to be an area of interest as society struggles with resistance to conventional antibiotics. It makes sense that moving forward there will be considerable interest in being able to predict bacterial susceptibility to 'treatment' phages based on sequence information alone. Furthermore, the next step would be modification of specific phages to improve their targeting/activity. This will rely on understanding of the phage genes that govern the specificity of infection in different backgrounds. The place to start is with certain key bacterial pathogens and a bank of phages.

2.5 Conclusions

In this chapter, the STEC O157 typing phages were clustered into four distinct groups of similar genomic sequences that broadly correlated with PT profile groups determined by two-way Euclidian clustering. Genetic variation within the TP groups may explain the subtle differences between the phage typing profiles exhibited by the STEC O157 TPs. This analysis was hindered by the lack of detailed annotation of protein encoding genes in T4 and T7-like phages. The impact of the order of transcription of the blocks of genetic loci and the role of host-induced modification further confound the analysis. However, sequencing the TPs has enabled the identification of the variable genes within each group and to determine how these correspond to changes in PT. Subsequent parts of this thesis focus on the genes that appear to alter host-phage interactions and aim to identify bacterial genes that influence TP resistance and susceptibility using random mutagenesis approaches. In order to understand the best combination of strains and individual phages to work with for phage therapy, the network of interactions needs to be analysed. This information can also provide insight into how phage typing can potentially be simplified in the future. A better understanding of the genetic differences between bacterial PTs, and the possible differences in virulence factors, could help elucidate why different PTs occupy specific niches and are associated with different patient age groups and disease severity.

Chapter 3: Effect of the loss and acquisition of stx prophage on PT

3.1 Background

In the last 25 years, strain replacement has been observed in the UK population of STEC O157 with the replacement of the PT2 phenotype with the PT21/28 phenotype [6] (fig. 1.5). The same study established that the direct ancestor of the now dominant PT21/28 clone is a PT32 strain. The highly successful PT21/28 clone accounts for >30% of clinical cases in the UK. This is because it has the *stx2a* genotype, either alone or in combination with *stx2c*, which is associated with disease severity. The PT32 ancestor of the PT21/28 clone had *stx2c* but was *stx2a* negative. This chapter describes an investigation into whether the gain of *stx2a* was responsible for increase in clinical cases associated with this clone and the phage type conversion.

In this chapter, it was established that the removal of the Stx2a phage from a PT21/28 strain resulted in the conversion of the strain to PT32. To further establish the mechanism by which this happened, two closely related strains with the same Stx2a subtype, one of the PT21/28 phenotype and one of the PT32 phenotype, were compared genotypically. I analysed the variation on the Stx2a prophage region to determine the genes that were responsible for the switch in PT were analysed.

3.2 Methods

3.2.1 stx lysogen deletion from STEC O157 strains

The lysogen curing was conducted by Sean McAteer in the laboratory of Prof. David Gally.

Lysogen curing is a naturally-occurring process in every lambdoid lysogen population. It results in loss of cI, the gene required for repression of pL. This loss of repression results in expression of tetracycline resistance (*Tc^R*) from the pTOF construct. The results in selection of the lysogen-cured strain from the population. The pTOF construct is thermosensitive (ts), which allows for it to be eliminated from the lysogen-cured strain.

The tetracycline resistance gene (*Tc^R*) CDS and native ribosome binding site without promoter from pBR322[116] were amplified by PCR using the Nt-PstI/-BglII and Ct-XhoII primers:

```
Nt-pTOF24-TcR
aaa ctgcag agatct taacgcagtcaggcaccgtgtatg
Ct-pTOF24-TcR
aaa ctcgag cgaggtgccgccggcttccattca
```

The fragment was cloned into pTOF24 [117], and colonies exhibiting chloramphenicol resistance (*Cm^R*) and screen kanamycin sensitive (*Km^S*) and tetracycline sensitive (*Tc^S*) were selected and designated pTOF-Tc^R.

The O_L (operator for leftward transcription, binding sites for CI and Cro repressors) and P_L (left promotor) were amplified by PCR from the STEC O157 lysogen using the 5'-PstI and 3'BGlII primers:

```
Stx2c phage:
5'-Sp5pL-PstI-IF
gtctcggtacccgac ctgcag cctctcgcccaaaaaaacacataac
3'-Sp5pL-BglII-IF
```

```
tgcctgactgcgtta agatct tgccagtctgttccatttggcttcc

Stx2a phage:
5'-560stx2pL-PstI-IF
gtctcggtacccgac ctgcag ctttgcctcacgttcgcccacc
3'-560stx2pL-BglII-IF
tgcctgactgcgtta agatct tcctgctgacgatgataataatg
```

The fragment was cloned into pTOF24-Tc^R and colonies exhibiting Cm^R and Tc^R to give pTOF24-O$_{L/}$P$_L$-Tc^R were selected. pTOF24-O$_{L/}$P$_L$-Tc^R was transformed into the lysogen and colonies exhibiting Cm^R were selected. Tc^R was confirmed to be lower in lysogen than non-lysogen and sponstaneous lysogen-cured with Tc^R was then selected and purified. Lysogen curing was verified by junction PCRs andpTOF24-O$_{L/}$P$_L$-Tc^R was removed from spontaneous lysogen-cured colonies.

3.2.2 Stx2a prophage comparison on 16438 and 9000

Long read PacBio sequenced complete genomes of strain 16438 (PT32 Stx 2a/2c) and strain 9000 (PT21/28 Stx 2a/2c) were compared to investigate the prophage Stx2a region of the genome. The Stx2a prophage was inserted in the same place and only varied by a few genes. These were compared in the genome comparison programs Mauve [118] and sequence identity between the two strains was compared by BLAST [119].

3.2.3 CI cloning

16438

```
TTTCTTAAGATTTCCAATAGTGAATAGTTAGTTGAAAGGTATGCGTGGAAACGCATGTAGCCTTAG
TTGGTCAGATATATTGGGACTCGCTTTGTCAGCGACGTAGGACGAATGTCCATTGTGAAAATAGCG
GTGTTACTTATGCAGCCGATGCTCTACGCGATACGAACACTAGGTTTTCCTTTTTCACAGGTTTAT
AACCCGTGAAATTACGAGTAGCTTCTTCGATTGCATTCGCTTTATCAGGGGAAGCTCTTCGAAATC
CATATGCAATCTGGTCAAGATAGCCAACTGAAGTTTTCGCTAATGCGGCGAGTCGCTTCCATTCCT
CACTAGAAGCCTCTTTTCGCCAGCGTAGTAGTTCATTACTCATTAGTGCCTCCGTTTATCACACAG
AATAACTTTACCATTTTGATAAATCAACCGCAATGTAAATTTATCATATTGCGTATTTATCCATTT
GCTAAATAGAGGGAAAATTGTGAGATGGAAAACAAAGATATTCGCAAATCGAATCTGGCGTTTTTG
CTAGATGAGCATAAAAAAATCGCGGGTAACACTAATGCAAGCTTTGCCGATAAGCTTGGGGTTAGC
CCTTCTCAACTCACGCAAGTCTCCGGTGAAAAAAGCACTCGAAACATAGGGGATAAACTAGCAAGA
AAATTTGAAGCCGCGCTTGGGTTACCTAATGGGTGGCTTGATTTGGTACATGATGTAACACCAATT
GCATCATGCTCAGATTCTTTAACTTTTGTCGGTCAGGTAAGAAAAGGGTTAGTGCGCGTGGTTGGT
GAGGCAATTCTTGGTGTTGATGGTGCCATCGAGATGACCGAAGAGCGCGATGGGTGGCTCAAAATT
TATAGCGATGATCCAGATGCCTTTGGTCTTCGTGTGAAAGGAGACAGCATGTGGCCCAGAATAAAA
TCAGGAGAATATGTACTCATTGAGCCTAACACCAAAGTATTCCCGGGTGATGAGGTGTTTGTCAGA
ACCGTTGAAGGACACAACATGATTAAGGTTCTTGGCTATGACAGAGATGGAGAATACCAATTTACA
AGCATTAACCAGGATCACAGGCCTATAACGTTGCCTTATCATCAAGTAGCAAAGGTGGAGTATGTG
GCTGGTATTCTGAAGCAATCTCGCCATCTGGATGACATCGAGGCAAGGGAGTGGCTGAAAAGTTCG
TGACTTCATCGTCACATAGCTGATAGCCAGTGGCCAGAAGAGACGTTTGGGTGATGAAACCACTTT
TATCTACAATTTACAGGGCGGTAAACATTGGCAAAAATAGATGATTATCAGCCAAGCCAAGTAGAA
GTTGATAAAGTACTTTATTGTAAAAAAATAGTTAACTTTTCTGGCGTTAAATGGAAACAGAAACCA
AGTCG

cro CDS

cI CDS

Cro sequence
MSNELLRWRKEASSEEWKRLAALAKTSVGYLDQIAYGFRRASPDKANAIEEATRNFTGYKPVKKEN
LVFVSRRASAA

CI sequence
MENKDIRKSNLAFLLDEHKKIAGNTNASFADKLGVSPSQLTQVSGEKSTRNIGDKLARKFEAALGL
PNGWLDLVHDVTPIASCSDSLTFVGQVRKGLVRVVGEAILGVDGAIEMTEERDGWLKIYSDDPDAF
GLRVKGDSMWPRIKSGEY
VLIEPNTKVFPGDEVFVRTVEGHNMIKVLGYDRDGEYQFTSINQDHRPITLPYHQVAKVEYVAGIL
KQSRHLDDIEAREWLKSS
```

Figure 3.1 Sequence extraction from the 16438 (PT32) genome containing the cro coding DNA sequence (CDS) in red and the cI CDS in blue. The amino acid sequence of each CDS is translated from the DNA sequence and listed below.

The cI CDS (blue region in fig. 3.1) was amplified from 16438 using primers listed in table 3.1.

Table 3.1 A table listing the nucleotide sequence of primers to clone cI and cro from strain 16438 into the cloning vectors pACYC184 and pWKS30.

Primers for cloning cI CDS plus pRM into pACYC184	
Nt-557stx2a/cI-pACYC184	Ct-557stx2a/cI
aaaaa gtcgac TTAGTGCCTCCGTTTATCACAC	aaaaa ggatcc ATCAGCTATGTGACGATGAAG
Primers for cloning cI CDS plus rbs into pWKS30	
Nt-557stx2a/cI-pWKS30	Ct-557stx2a/cI
aaaaa gtcgac AAATAGAGGGAAAATTGTGAG	aaaaa ggatcc ATCAGCTATGTGACGATGAAG
Primers for cloning cro-cI into pACYC184	
Ct-560stx2a/cro-pACYC184	Ct-557stx2a/cI
aaaaa ggatcc TTGTGAAAATAGCGGTGTTAC	aaaaa ggatcc ATCAGCTATGTGACGATGAAG

Cloned products were confirmed with sanger sequencing. The fragments were cloned into pACYC184 [120] and Tc^R were selected to give pACYC184-16438-pRM-CI. TcR was cloned into pWKS30 [121] and Ap^R was selected to give pWKS30-16438-rbs-CI. This was transformed into strain 9000 by electroporation and Tc^R was selected for pACYC184-16438-pRM-CI and Ap^R for pWKS30-16438-rbs-CI . Susceptibility to typing phages 6 and 13 was tested.

9000

```
TTTCTTAAGATTTCCAATAGTGAATAGTTAGTTGAAAGGTATGCGTGGAAACGCATATGGCCTTAG
TTGGTCAGATATATTGGGACTCGCTTTGTCAGCGACGTAGGACGAATGTCCATTGTGAAAATAGCG
GTGTTACTTATGCAGCCAGAAGGTTCTTTTTGCTTATTTCAAGCATTTCGCTTGCTTGATATTTGC
CACCAGAAATCTCTTCGATTTTTGATGCGTATTTAGTTTTCCCAAAAAACTCAGTCTTAGGGAGGA
AGCCGTTTTTGAGCCACTTATAGACAGCCCTTTCGCTAACTCCACAAGCCTTCGCAACTTCAGGGA
TGCCGACACCTTTAATCGGCTCATCAAGATTTTGCATAGGAATATCCTTTTTCGTACTTTCAGTAC
GTATTATGGTTGAACTGAAAGTTTTTGCAAGTGCTTTAGTATCGTACTCATGGTTCAGAATGAAAA
AGTGCGCAAAGAATTCGCCCAGCGGCTAGCGCAAGCCTGTAAAGAAGCTGGTCTTGATGAACATGG
TAGGGGAATGGCTATAGCCCGTGCCCTTTCTCTTTCGTCCAAAGGCGTTAGCAAATGGTTTAATGC
TGAGTCTTTACCGCGTCAGGAAAAAATGAATGCGCTTGCGAAATTTCTAAACGTTGATGTTGTTTG
GCTTCAGCACGGCACTTCGTTAAATGGAGCGAATGATGAAGATACTCTTTCATTTGTTGGCAAATT
AAAAAAAGGGTTAGTGCGCGTGGTTGGTGAGGCAATTCTTGGTGTTGATGGTGCCATCGAGATGAC
CGAAGAGCGCGATGGGTGGCTCAAAATTTATAGCGATGATCCAGATGCCTTTGGTCTTCGTGTGAA
AGGAGACAGCATGTGGCCCAGAATAAAATCAGGAGAATATGTACTCATTGAGCCTAACACCAAAGT
ATTCCCGGGTGATGAGGTGTTTGTCAGAACCGTTGAAGGACACAACATGATTAAGGTTCTTGGCTA
TGACAGAGATGGAGAATACCAATTTACAAGCATTAACCAGGATCACAGGCCTATAACGTTGCCTTA
TCATCAAGTAGCAAAGGTGGAGTATGTAGCTGGTATTCTGAAGCAATCTCGCCATCTGGATGACAT
CGAGGCAAGGGAGTGGCTGAAAAGTTCGTGACTTCATCGTCACATAGCTGGTAACCAGTGGCCTGA
AGAGACGTTTGGGTAAGGAGGATAGATGGCGTTCAATGACCTTGAATATCAAGCAGTAAAAAAAGA
AGTTCACCAATTCATTGAAAGCATAAGGCCGCCTGAACATATCCGCAATGAACTGGATATTGTTTA
TAGCATCAATGACCAAACGATAGATATCGGCGA

cro CDS

cI CDS

Cro sequence
MQNLDEPIKGVGIPEVAKACGVSERAVYKWLKNGFLPKTEFFGKTKYASKIEEISGGKYQASEMLE
ISKKNLLAA

CI sequence
MVQNEKVRKEFAQRLAQACKEAGLDEHGRGMAIARALSLSSKGVSKWFNAESLPRQEKMNALAKFL
NVDVVWLQHGTSLNGANDEDTLSFVGKLKKGLVRVVGEAILGVDGAIEMTEERDGWLKIYSDDPDA
FGLRVKGDSMWPRIKSGEYVLIEPNTKVFPGDEVFVRTVEGHNMIKVLGYDRDGEYQFTSINQDHR
PITLPYHQVAKVEYVAGILKQSRHLDDIEAREWLKSS
```

Figure 3.2 Sequence extraction from the 9000 (PT21/28) genome containing the cro coding DNA sequence (CDS) in red and the cI CDS in blue. The amino acid sequence of each CDS is translated from the DNA sequence and listed below.

The cI CDS (blue region in fig. 3.2) were amplified from 9000 using primers listed in table 3.2.

Table 3.2 A table listing the nucleotide sequence of primers to clone cI and cro from strain 9000 into the cloning vectors pACYC184 and pWKS30.

Primers for cloning cI CDS plus pRM into pACYC184	
Nt-560stx2a/cI-pACYC184	Ct-560stx2a/cI
aaaaa gtcgac TAGGAATATCCTTTTTCGTAC	aaaaa ggatcc ACCAGCTATGTGACGATGAAG
Primers for cloning cI CDS plus rbs into pWKS30	
Nt-560stx2a/cI-pWKS30	Ct-560stx2a/cI
aaaaa gtcgac AAGTGCTTTAGTATCGTACTC	aaaaa ggatcc ACCAGCTATGTGACGATGAAG
Primers for cloning cro-cI into pACYC184	
Ct-560stx2a/cro-pACYC184	Ct-560stx2a/cI
aaaaa ggatcc TTGTGAAAATAGCGGTGTTAC	aaaaa ggatcc ACCAGCTATGTGACGATGAAG

Cloned products were confirmed with sanger sequencing. The fragments were cloned into pACYC184 [120] and Tc^R were selected to give pACYC184-9000-pRM-CI. TcR was cloned into pWKS30 [121] and Ap^R was selected to give pWKS30-9000-rbs-CI. This was transformed into strain 16438 by electroporation and Tc^R was selected for pACYC184-9000-pRM-CI and Ap^R for pWKS30-9000-rbs-CI . Susceptibility to typing phages 6 and 13 was tested.

3.2.4 Biolog Phenotyping Microarray

A single isolated colony of strains 9000 and 1460 (table 3.3) was grown in difco broth overnight. Into phenotyping microarray (PM) plates 1 (Biolog, Inc., Hayward, CA), 100 ul/well was added. The plates were incubated at 33°C for 48 hours using the Omnilog II Combo System (Biolog, Inc., Hayward, CA). The output data from the Omnilog was imported into the opm package in R for analysis [122].

3.3 Results

3.3.1 Stx2a deletion from PT21/28 strain converts it to become PT32

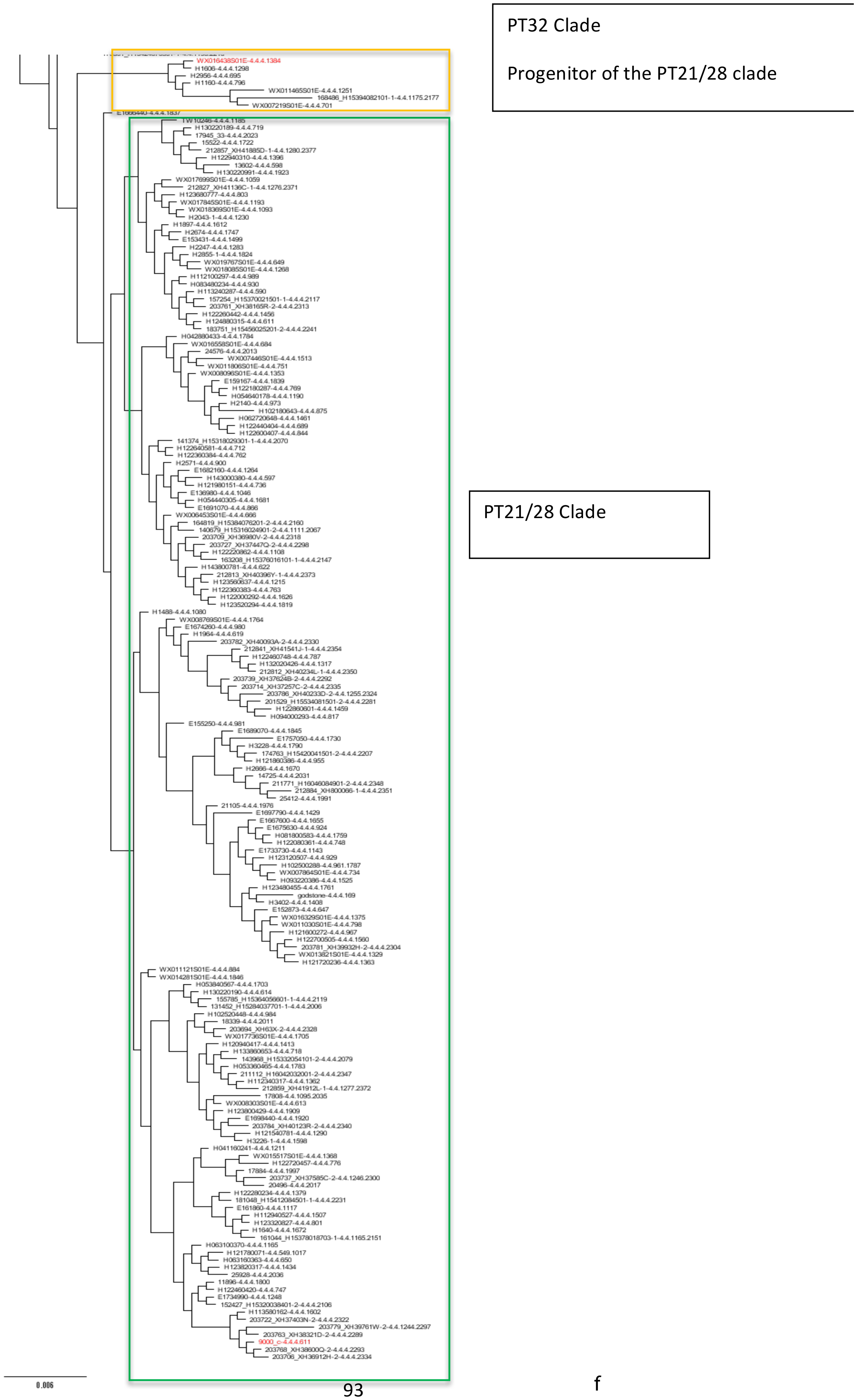
Lysogen curing of strains 9000 and 10671 from the UK IPRAVE study [123] for both the Stx2a and Stx2c, individually and together, revealed the loss of Stx2a changed the bacteriophage susceptibility phenotype from PT21/28 to PT32. The lysogen curing effect on PT is listed in detail in table 3.3. It is clear from these results that the loss of Stx2c does not affect the PT21/28 phenotype but that the loss of Stx2a as a whole phage results in the switch in PT. It is also evident that the loss of the whole phage is required and not the toxins alone as was shown by 1599 remaining PT21/28.

Table 3.3 A table that lists the PT and Stx content of original and lysogen cured versions of strains 9000 and 10671 and their observed PT phenotype.

Roslin Ref	Description	Phage type
9000	Original PT21/28 IPRAVE isolate, stx2a & stx2c	21/28
10671	Original PT32 IPRAVE isolate, stx2c only	32
1456	9000 with stx2c phage partly deleted	21/28
1460	9000 with stx2a phage entirely deleted	32
1463	10671 with stx2c phage entirely deleted	32
1465	1456 with stx2a phage entirely deleted	32
1467	1460 with stx2c phage partly deleted	32
1599	9000 with stx2a & stx2c toxin genes entirely deleted	21/28

From these results, it was evident that a component of the Stx2a phage was responsible for the resistance to typing phages 6 and 13 observed in the PT21/28 phenotype but not in the PT32 phenotype. Chapter 2 established that phages 6 and 13 are both members of the group 2 typing phages and are very closely related. This means that the resistance to one is likely to also confer resistance to the other and it may be caused by just one gene. These results also support the hypothesis that the PT21/28 clade emerged from the PT32 clade of STEC O157 with the acquisition of Stx2a [6].

A clade of strains belonging to PT32, identified as the progenitor strains of PT21/28 (fig. 3.3), had acquired the stx2a-encoding phage but had retained the PT32 pattern – ie. susceptibility to TP 6 and 13. A member of this clade (16438) was chosen to compare to 9000 to see how genotypically similar the Stx2a region was in the two strains, the hypothesis being that one or some of the genes that varied would be responsible for the change in PT.



PT32 Clade

Progenitor of the PT21/28 clade

PT21/28 Clade

Figure 3.3 Section of a phylogenetic tree of STEC O157 showing the evolution of the the PT21/28 UK prevalent clade from the PT32 clade. Strains 16438 and 9000 are highlighted in red.

3.3.2 High level of similarity between strain 16438 and 9000 Stx2a prophage regions

The Stx2a prophage region is 50kbp and was inserted in the same place in the two strains. A blast comparison of the 16438 Stx2a prophage region and the 9000 Stx2a prophage region revealed a 99% identity at a 93% coverage level. Further analysis of the regions that differed in Mauve revealed that there were several genes that were unique to each phage and that there were changes in insertion element positioning. The changes are detailed in figure 3.4.

Two transposase mediated IS regions had changed position between the two strains but this is not likely to have effected PT. More interestingly there were twelve genes that were found in one Stx2a phage but not the other. Five of these were hypothetical proteins but of note there was a completely unique serine/threonine kinase only found on the strain 9000 stx2a region.

The N terminal variation in the CI repressor was of interest because bacteriophage use CI to maintain lysogeny. During lysogeny CI is the only protein being expressed by a prophage that represses transcription and induction of the bacteriophage [124]. The presence of CI also causes immunity to superinfection by other lambda phages [125]. It was therefore necessary to test whether the constitutively expressed CI from stx2a in 9000 repressed lytic phage activity, indicating that the CI would be involved in immunity from lytic and lysogenic infection.

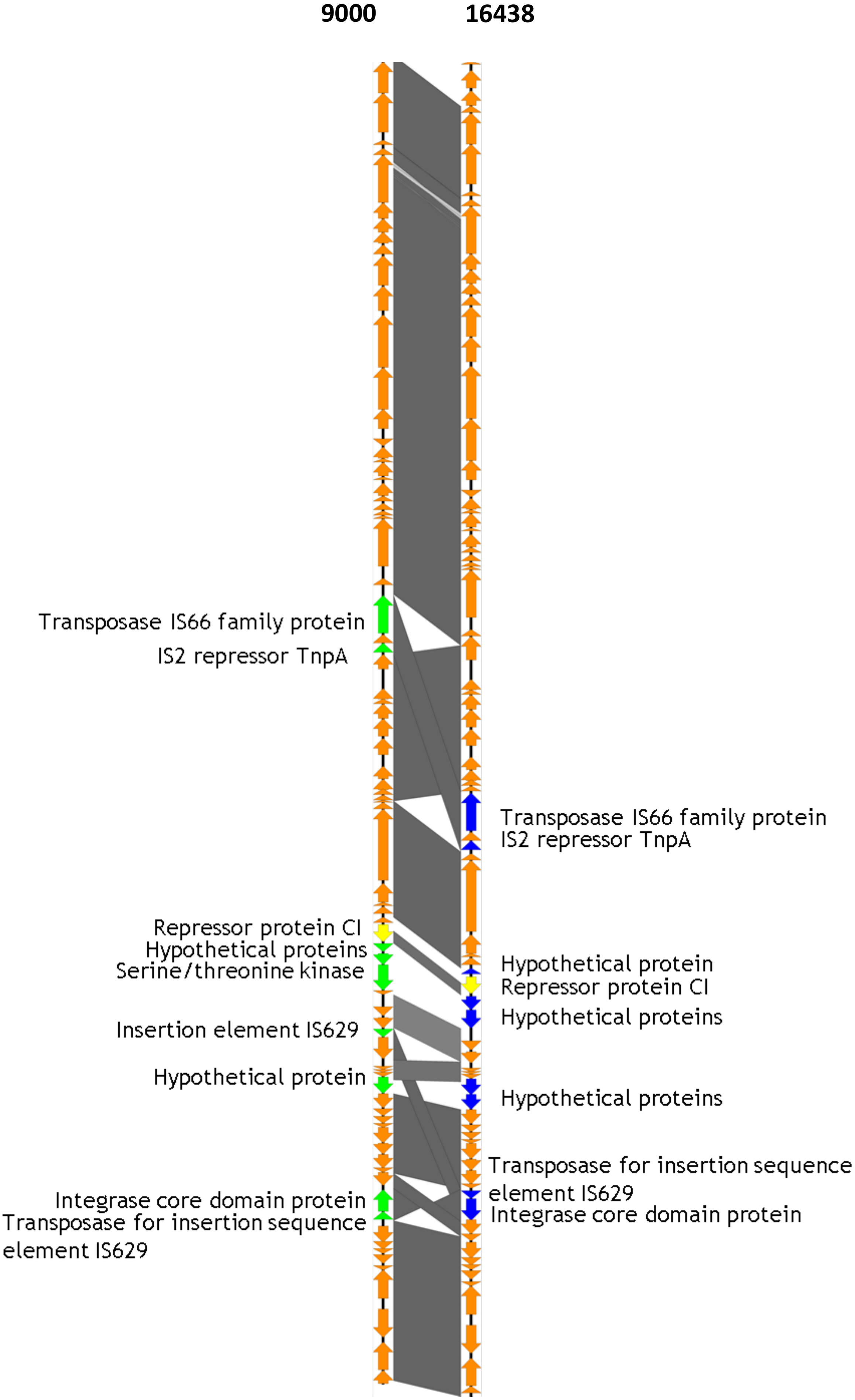

9000
16438
Transposase IS66 family protein
IS2 repressor TnpA
Transposase IS66 family protein
IS2 repressor TnpA
Repressor protein CI
Hypothetical proteins
Serine/threonine kinase
Hypothetical protein
Repressor protein CI
Hypothetical proteins
Insertion element IS629
Hypothetical protein
Hypothetical proteins
Transposase for insertion sequence element IS629
Integrase core domain protein
Integrase core domain protein
Transposase for insertion sequence element IS629

Figure 3.4 A schematic diagram of the regions that vary between the Stx2a prophage region in strain 9000 and strain 16438 drawn with easyfig [106]. The arrows represent the genes on each genome and the blast hits between the two strains are represented by the blocks of grey between them. Orange arrows are the genes that are the same in both strains, green are genes that are different in strain 9000, blue is genes that are different in strain 16438 and yellow is used to mark the CI repressor protein that varies at the N terminal end between the two strains.

3.3.3 CI cloning

CI effect on phage susceptibility was tested by cloning the CI CDS from strain 16438 into strain 9000 and the CI CDS from strain 9000 into 16438. The constructs were then tested with typing phages 6 and 13 to look for a change from WT susceptibility to construct resistant and vice versa. Unfortunately, neither cloning vector plasmid (pACYC184 and pWKS30) transformation into strains 9000 or 16438 changed the phage type susceptibility from the WT.

3.3.4 Change in substrate utilization of Stx2a knockouts

To test whether the Stx2a phage has any effect on the metabolism of the host, a selection of substrates were assayed using the biolog phenotyping arrays. The Stx2a knockout had showed increased utilization in 31 substrates compared to the WT (table 3.4). These included Dextrin, D-Trehalose, α-D-glucose, D-Mannose, D-Fructose, Sucrose, Inosine, D-Raffinose, α-D-Lactose, D-Melibiose, ß-Methyl-D-Glucoside, N-Acetyl-D-Galactosamine, N-Acetyl Neuraminic Acid, Glycerol, D-Glucose-6-PO4(Hexose), D-fructose-6-PO4, Methyl Pyruvate, Glycyl-L-Proline, L-Lactic Acid, L-Alanine, L-Aspartic acid, D-Malic acid, L-Serine, D-Galaturonic Acid, L-Galacturonic Acid Lactone, D-Gluconic acid, D-Glucoronic acid, Glucuronamide, Proprionic Acid, Acetic Acid and Aztreonam. It also showed a large decrease in D-Serine utilization.

Table 3.4 A table that details the results of the Biolog phenotyping array from the Stx2a knockout 1460 compared to the WT 9000 on a list of metabolic substrates. Those substrates that have an increased utilization with a change in growth rate >10 are highlighted in yellow. The $\log_2$ rate of formazan production, measured as log2[(A_{590}-A_{750})/hr] is shown in backets. The decrease in D-Serine utilization is highlighted in green.

Substrate	**Utilization of Stx2a knockout compared to WT** **(Growth rate of WT and knockout)**	**Substrate**	**Utilization of Stx2a knockout compared to WT** **(Growth rate of WT and knockout)**
Negative control	None (68 and 71)	1% NaCl	None (185 and 175)
Dextrin	Increased (138 and 151)	4% NaCl	None (82 and 86)
D-Maltose	None (86 and 93)	8% NaCl	None (98 and 94)
D-Trehalose	Increased (113 and 138)	α-D-glucose	Increased (93 and 106)
D-Cellobiose	None (70 and 71)	D-Mannose	Increased (130 and 145)
Gentiobiose	None (70 and 74)	D-Fructose	Increased (136 and 164)
Sucrose	Increased (90 and 108)	D-Galactose	None (194 and 202)

D-Turanose	None (68 and 73)	3-Methyl Glucose	None (44 and 46)
Stachyose	None (72 and 73)	D-Fucose	None (75 and 81)
Positive control	None (162 and 158)	L-Fucose	None (215 and 223)
pH 6	None (122 and 120)	L-Rhamnose	Increased (72 and 76)
pH 5	None (111 and 115)	Inosine	Increased (145 and 173)
D-Raffinose	Increased (94 and 115)	1% Sodium Lactate	None (224 and 219)
α-D-Lactose	Increased (103 and 118)	Fusidic acid	None (165 and 158)
D-Melibiose	Increased (140 and 164)	D-Serine	None (85 and 89)
β-Methyl-D-Glucoside	Increased (54 and 67)	D-Sorbitol	None (57 and 61)
D-Salicin	None (51 and 56)	D-Mannitol	None (73 and 83)
N-Acetyl-D-Glucosamine	None (104 and 113)	D-Arabitol	None (49 and 52)
N-Acetyl-β-D-Mannosamine	None (64 and 72)	myo-Inositol	None (39 and 41)
N-Acetyl-D-Galactosamine	Increased (140 and 154)	Glycerol	Increased (190 and 203)
N-Acetyl Neuraminic Acid	Increased (118 and 138)	D-Glucose-6-PO4(Hexose)	Increased (209 and 243)
Quinic Acid	None (57 and 58)	D-fructose-6-PO4	Increased (229 and 251)
D-Saccharic Acid	None (62 and 67)	D-aspartic acid	None (56 and 59)

Vancomycin	None (268 and 266)	D-Serine	Decreased (155 and 57)
Tetrazelium Violet	None (308 and 303)	Troleandemycin	None (237 and 229)
Tetrazelium Blue	None (253 and 243)	Rifamycin SV	None (214 and 203)
p-Hydroxy-Phenylacetic Acid	None (50 and 53)	Minocycline	None (88 and 91)
Methyl Pyruvate	Increased (66 and 81)	Gelatin	None (62 and 65)
D-Lactic Acid Methyl Ester	None (45 and 46)	Glycyl-L-Proline	Increased (91 and 112)
L-Lactic Acid	Increased (205 and 224)	L-Alanine	Increased (98 and 126)
Citric Acid	None (38 and 43)	L-Arginine	None (41 and 46)
α-keto-Glutaric Acid	None (47 and 50)	L-Aspartic acid	Increased (68 and 90)
D-Malic acid	Increased (96 and 123)	L-Glutamic acid	None (53 and 58)
L-Malic acid	None (87 and 94)	L-Histidine	None (45 and 50)
Bromo-Succinic Acid	None (71 and 76)	L-Pyroglutamic acid	None (56 and 60)
Nalidixic Acid	None (70 and 77)	L-Serine	Increased (173 and 189)
Lithium Chloride	None (193 and 201)	Lincomycin	None (269 and 265)
Potassium Tellurite	None (246 and 249)	Guanidine HCl	None (225 and 219)

Tween-40	None (58 and 59)	Niaproof 4	None (224 and 212)
γ-Amino-Butyric Acid	None (37 and 45)	Pectin	None (51 and 48)
α-Hydroxy-Butyric Acid	None (54 and 64)	D-Galaturonic Acid	Increased (184 and 202)
β-Hydroxy-D,L-Butyric Acid	None (50 and 52)	L-Galacturonic Acid Lactone	Increased (208 and 229)
α-Keto-Butyric Acid	None (49 and 59)	D-Gluconic acid	Increased (152 and 195)
Acetoacetic Acid	None (80 and 80)	D-Glucoronic acid	Increased (213 and 231)
Proprionic Acid	Increased (85 and 99)	Glucuronamide	Increased (169 and 181)
Acetic Acid	Increased (120 and 140)	Mucic Acid (mucose)	None (62 and 70)
Formic Acid	None (60 and 65)	Sodium Butyrate	None (243 and 248)
Aztreonam	Increased (111 and 202)	Sodium Bromate	None (114 and 106)

3.4 Discussion

The PT21/28 phenotype is a UK specific clone that emerged suddenly around 25 years ago and outcompeted PT2 to become the dominant STEC O157 clone in the UK. This clone is associated with more severe disease and it poses a significant public health threat and is responsible for hundreds of domestically acquired STEC O157 cases every year. It is likely that the emergence of this highly pathogenic PT21/28 clone is linked to the high incidence of STEC O157 in the UK in relation to other countries. The work in this chapter showed that this clone emerged when the PT32 ancestor acquired Stx2a. It is likely that the clone rapidly became successful due to increased pathogenicity, bacteriophage resistance and metabolic adaptation.

The Stx encoding prophage region is an important marker for clinical severity but can also be used as a marker for determining the environment or geographical region that the strain has come from. Knowing the geographical source of the strain is highly important for epidemiological tracking of outbreaks and certain Stx phages may be more associated with certain environments or regions [126]. Ruminants are almost invariably the original source of STEC O157 but variants of the prophage family seen in different strains may reveal the evolutionary history of a strain and the source or environment of origin. It is clear that strains 16438 and 9000 have come from a similar environment and are closely related with respect to both their core and accessory genome, including their prophage content. However, the two strains have undergone separate selective pressures that have resulted in varying Stx2a prophages. This has resulted in a phenotypic difference in the lytic phage susceptibilities of these strains (susceptibility to typing phages 6 and 13) and the resulting phage type.

This analysis showed that the presence of the variant CI alone was not sufficient to change the PT but it was hypothesized that the CI may also need Cro to be plasmid-encoded. Even though the Cro region is the same in both strains, it might be that the CI requires Cro to be present on the same expression vector and expressed in tandem. This would mean that any interactions they may have that could be causing the change in PT would be more likely to happen. It will be interesting to see if Cro has any effect on PT in future studies. It may also be required that the CI is provided on the full bacteriophage background and not as single gene clones for the same observed phenotypic change. It is known that the

Cro and CI repressors together form a helix-turn-helix superfamily [127] and Cro is used to turn off early gene transcription during the lytic cycle [128]. The two repressors may be in communication with each other to perform immunity to both lytic and lysogenic phages in partnership.

The unique genes found on Stx2a prophage region in 16438 could be associated with the infectivity of typing phages 6 and 13 but it is far more likely that the unique genes found on the Stx2a prophage region in 9000 are responsible for the increased bacteriophage resistance seen in the PT21/28 strain. This is because the new prophage genes are more likely to confer superinfection immunity by some mechanism that has been documented in numerous studies[125, 129] but prophage genes have not yet been implicated in susceptibility. Apart from CI and the Serine/threonine kinase, the other unique genes are all annotated as hypothetical proteins so it was difficult to speculate what involvement they might have.

The unique Serine/Threonine kinase was of interest because the presence of Stx2 is known to have an effect on the metabolism of the host [130]. The presence of Stx2a has been shown to metabolically restrict the host for 31 substrates as the knockout strain showed increased substrate utilization for these 31 compared to the WT (table 3.4). Notably, the Stx2a knockouts of strain 9000 were shown to have decreased D-Serine utilization compared to the WT. Although the unique gene found on strain 9000 is annotated as a Serine/threonine kinase it may actually play a wider role in substrate utilization restriction and be associated with pathoadaptivity as well as potentially being involved in lytic bacteriophage resistance. Loss of metabolic activity may have resulted in this clone becoming pathoadapted to the bovine and/or human host, indicating that the gain of Stx2a provides multiple advantages to the STEC O157 in increased resistance to certain bacteriophages, pathoadaptivity and increased pathogenicity.

The evolution of the PT21/28 clade from the PT32 ancestor by the addition of Stx2a and its subsequent success in terms of clinical prevalence, is evidence of the increasing pathogenicity of STEC O157 and the danger that it poses. The potential for Stx negative *E. coli* strains to gain a Stx prophage is high in certain environments [131]. This also applies to the gain of additional Stx prophage regions in addition to those present in stains of STEC O157. There is a diversity of free Shiga toxin encoding bacteriophages that have been found in sewage in multiple countries [132] and environments contaminated with cattle faeces [133]. The work in this chapter has indicated that not only does the gain of a Stx2a phage provide pathogenic potential to a strain of STEC O157 but it also provides immunity to some bacteriophages and therefore may confer an evolutionary advantage.

Interestingly the difference between PT1 and PT8 is also the gain in resistance to TP 6 and 13 in PT1. So it is possible that they might also have gained this resistance by a similar mechanism to the PT32 > PT21/28 switch. This could be tested by comparing the Stx2a prophage regions in closely related PT1 and PT8 strains, in a similar way to the analysis done in this chapter. It would be interesting to see if any unique genes found on a Stx2a prophage region in PT1 were the same as those observed in strain 9000.

3.5 Conclusions

The work described in this chapter demonstrates that the deletion of the Stx2a prophage region from a PT21/28 strain converted it to PT32. Bioinformatic analysis revealed that both the PT32 and PT21/28 strains had very similar Stx2a prophage regions that were inserted into the same place. The Stx2a prophages only varied by a limited number of genes so it was concluded that one or a combination of these unique genes were responsible for the altered phenotype. The initial hypothesis that the presence of the CI repressor had an impact on phage susceptibility was not proven. More experimental cloning work is required to investigate this further. It was proposed that the unique Serine/threonine kinase had the potential to alter the phage susceptibility phenotype. However, this work has provided valuable evidence that the highly successful PT21/28 clone that accounts for >30% of clinical cases in the UK evolved from a PT32 ancestor with the acquisition of the Stx2a phage. The prophage region conferred an advantage to the strain in terms of bacteriophage resistance, enhancing the chances of survival in the bacteriophage colonized human and bovine guts. This clone has expanded rapidly since the acquisition of Stx2a and is now the most clinically significant domestic strain of STEC O157 in the UK today.

Chapter 4: Impact of short and long read sequencing technologies in understanding PT differences

4.1 Background

The initial hypothesis for the project was that analysis of whole genome short read sequences of ~250 strains of STEC O157 would highlight genetic variation that would correlate with phage type and elucidate the molecular mechanisms of phage type/ host interactions. It became clear with the use of short read sequencing that the genetic relatedness of PT was a complicated issue and that it may not always be the same mechanism causing the same PT. It also became clear that to study mobile elements that could be influencing PTs long read sequencing technology would be required in order to provide better assembly of highly repetitive regions like plasmids or prophages. In this chapter, long read sequencing was used to investigate two closely related outbreaks that involved two associated strains of different PT to identify genetic variation responsible for the change in PT.

4.2 Methods

4.2.1 Selection of a range of STEC O157 strains of various PT to analyse with short read sequencing

20 outbreak and non-outbreak strains of varying phage type were selected (fig. 4.1), to try to obtain a cross-section of common PTs phylogenetically closely related as well as not, for closer analysis. Two PT 14s, six PT 8s, three PT 54s, three PT 31s, two PT 21/28s, two PT 32s and and two PT 2s were chosen.

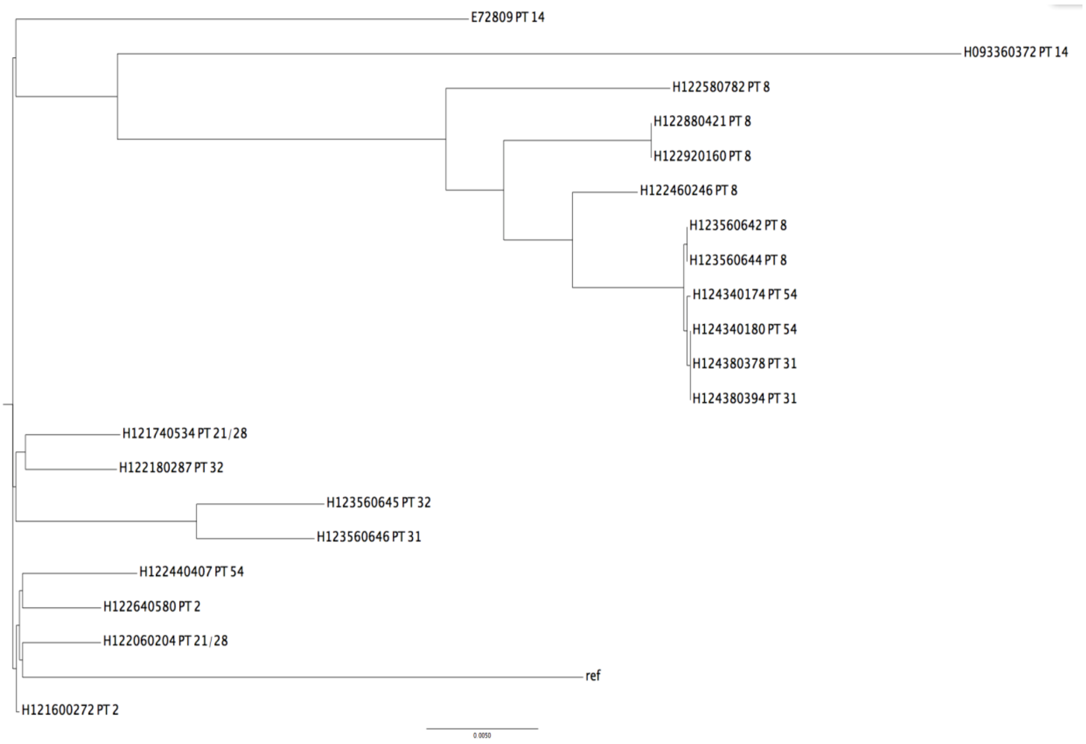
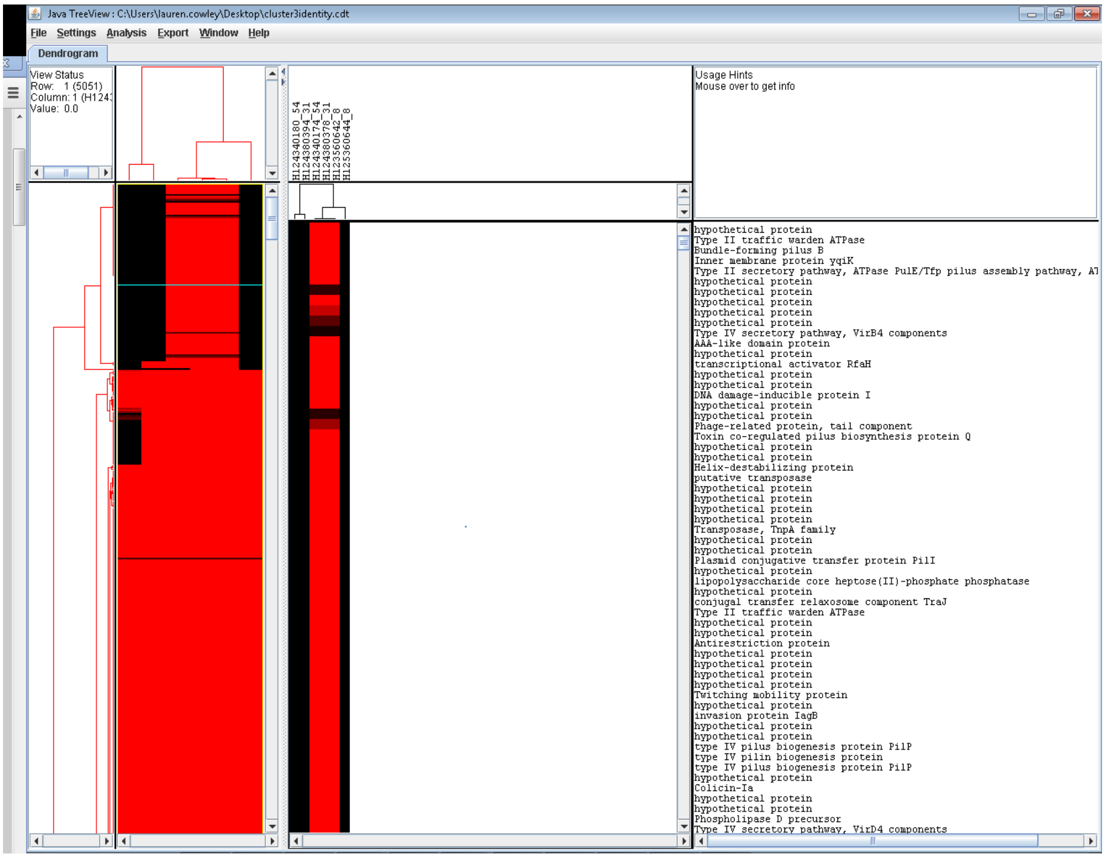


Figure 4.1 Maximum likelihood phylogenetic tree of strains selected for short read PT analysis with Sakai as a reference strain.

4.2.2 Heat-map comparison of short read sequences of various PTs

A heatmap of gene by gene homology was drawn to see if any correlations could be ascertained for different PTs. The strains were sequenced with Illumina technology, assembled with velvet [93] and annotated using prokka [94] from which the coding sequences were extracted. The identified genes were BLASTed [119] against each other to identify homologous genes and gene families were created using single linkage mode clustering at 85% sequence identity and >= 90% overlap with T-cluster. The heatmap was then viewed in Java TreeView [134] and gene differences were sought (fig. 4.2).



View Status
Row: 1 (5051)
Column: 1 (H1243
Value: 0.0

Usage Hints
Mouse over to get info

```
hypothetical protein
Type II traffic warden ATPase
Bundle-forming pilus B
Inner membrane protein yqiK
Type II secretory pathway, ATPase PulE/Tfp pilus assembly pathway, AT
hypothetical protein
hypothetical protein
hypothetical protein
hypothetical protein
hypothetical protein
Type IV secretory pathway, VirB4 components
AAA-like domain protein
hypothetical protein
transcriptional activator RfaH
hypothetical protein
hypothetical protein
DNA damage-inducible protein I
hypothetical protein
hypothetical protein
Phage-related protein, tail component
Toxin co-regulated pilus biosynthesis protein Q
hypothetical protein
hypothetical protein
Helix-destabilizing protein
putative transposase
hypothetical protein
hypothetical protein
hypothetical protein
hypothetical protein
Transposase, TnpA family
hypothetical protein
hypothetical protein
Plasmid conjugative transfer protein PilI
hypothetical protein
lipopolysaccharide core heptose(II)-phosphate phosphatase
hypothetical protein
conjugal transfer relaxosome component TraJ
Type II traffic warden ATPase
hypothetical protein
hypothetical protein
Antirestriction protein
hypothetical protein
hypothetical protein
hypothetical protein
hypothetical protein
Twitching mobility protein
hypothetical protein
invasion protein IagB
hypothetical protein
hypothetical protein
type IV pilus biogenesis protein PilP
type IV pilin biogenesis protein
type IV pilus biogenesis protein PilP
hypothetical protein
Colicin-Ia
hypothetical protein
hypothetical protein
Phospholipase D precursor
Type IV secretory pathway, VirD4 components
```

Figure 4.2 Screenshot of Java Treeview displayed heatmap that facilitates browsing through strains selected and observation of homology between shared proteins. Red indicates presence and black indicates absence of a gene.

4.2.3 Outbreak investigation

An investigation based on two associated outbreaks of STEC O157 with differing PTs from the same food outlet in 2012 was undertaken. In August 2012, four cases of STEC O157 PT8 were associated with visiting the food outlet. In October 2012, over 145 cases of STEC O157 PT54 all reported eating at the same restaurant.

4.2.4 Illumina sequencing and core SNP analysis

As part of the outbreak investigation eighty-nine isolates were selected for whole genome sequencing, including six from the August PT8 cluster and 53 from the PT54 cluster in October, and 30 isolates of STEC O157 from temporally and geographically related sporadic cases isolated between June and November 2012. Genomic DNA was fragmented and tagged for multiplexing with Nextera XT DNA Sample Preparation Kits (Illumina) and sequenced at the Animal Laboratories and Plant Health Agency using the Illumina GAII platform with 2x150bp reads. Short reads were quality trimmed [136] and mapped to the reference STEC O157 strain *Sakai* (Genbank accession BA000007) using BWA-SW [137]. The Sequence Alignment Map output from BWA was sorted and indexed to produce a Binary Alignment Map (BAM) using Samtools [137]. GATK2 [138] was used to create a Variant Call Format (VCF) file from each of the BAMs, which were further parsed to extract only single nucleotide polymorphism (SNP) positions which were of high quality (MQ>30, DP>10, GQ>30, Variant Ratio >0.9). Pseudosequences of polymorphic positions were used to create maximum likelihood trees using RaxML[139]. Pair-wise SNP distances between each pseudosequence were calculated. Spades version 2.5.1 [95] was run using careful mode with kmer sizes 21, 33, 55 and 77 to produce *de novo* assemblies of the sequenced paired-end fastq files. FASTQ sequences were deposited in the NCBI Short Read Archive under the BioProject PRJNA248042.

4.2.5 PacBio sequencing

One isolate of STEC O157 PT8 (ref 644-PT8) and one belonging to PT54 (ref 180-PT54) from the outbreak were selected. High molecular weight DNA was extracted using Qiagen Genomic-tip 100/G columns and a modification of the protocol previously described by Clawson et al. (2009)[140]. Ten micrograms of DNA was sheared to a targeted size of 20 kb using a g-TUBE (Corvaris, Woburn, MA) and concentrated using 0.45X volume of AMPure PB magnetic beads (Pacific Biosciences, Menlo Park, CA) following the manufacture's protocol. Sequencing libraries were created using 5 micrograms of sheared DNA and the PacBio DNA SMRTbell Template Prep Kit 1.0 and fragments 10 kb or larger selected using a BluePippin (Sage Science, Beverly, MA) with the smrtbell 15-20 kb setting. The library was bound with polymerase P5 followed by sequencing on a Pacific BioSciences (PacBio) RS II sequencing platform with chemistry C3 and the 120 minute data collection protocol.

A fastq file was generated from the PacBio reads using SMRTanalysis and error-corrected reads were created using PBcR with self-correction [141]. The longest 20X coverage of the corrected reads were assembled with Celera Assembler 8.1. The resulting contigs were polished using Quiver [142] and annotated using a local instance of Do-It-Yourself Annotator (DIYA)[143]. The annotated genome sequence was imported into Geneious (Biomatters LTD., Auckland, New Zealand) and duplicated sequence removed from the 5' and 3' ends to generate the circularized chromosome. The origin of replication was approximated using OriFinder [144] and the chromosome reoriented using the origin as base 1. PacBio sequenced strain 180-PT54 is available under accessions CP015832 for the chromosome and CP015833 for the plasmid.

4.2.6 MinION sequencing

DNA from isolate 644-PT8 was extracted using the STRATEC molecular invisorb spin minikit and diluted to a concentration of 1ug of genomic DNA in 50ul of water. The MinION library was prepared using the SQK-MAP006 genomic sequencing kit according to the manufacturer's instructions and sequencing was performed on a

Mk1 MinION with a Mk1 flow cell.

~76000 reads were produced and 26 fold passing 2D coverage of the genome was achieved. The long read assembly program Canu [141] was used to assemble the long reads and two chromosomal contigs were produced. The assembly of isolate 644-PT8 showed greater concordance with the synteny of isolate 180-PT54 than the assembly of the PacBio sequencing had produced, but was still not resolved in a similar region.

OpGen mapping for the strain was obtained from a commercial provider. The isolate 644-PT8 was rotated using OriFinder to have the same point of origin as isolate 180-PT54 strain for comparison. The genome was then annotated using PROKKA [94]. MinION sequenced strain 644-PT8 is available under accession CP015831.

4.2.7 Analysis of the accessory genome using long read sequences

The two annotated PacBio assemblies were analysed in PHAST [145] which identifies complete and incomplete prophage regions and their constituent genes. The gene annotations and their sequences for each strain were compared to each other with blastn [119]. Unique genes were extracted from the results if they had no match at greater than 70% nucleotide identity and overlap to any of the annotated genes in the other strain using a reciprocal blast approach.

NUCMER [146] was used to align the two assemblies and to identify single nucleotide polymorphisms, ambiguous alignment SNPs were excluded.

Gene annotations were extracted from the genbank file and resistance annotations were manually analysed for differences between the two strains.

Plasmid differences were visualised in Mauve [118]. The plasmid from isolate 180-PT54 was graphically visualised using BRIG [105] and compared to other IncHI2 plasmid sequences found in Genbank that were highly similar by blast at >98%

identity and >60% coverage (Accession numbers KM877269.1, JN983042.1, BX664015.1, DQ517526.1, EF382672.1, LN794248.1, LK056646.1, EU855787.1, KP975077.1, EU855788.1, CP011601.1, CP008906.1, CP008825.1, CP012170.1 and CP006056.1).

4.2.8 Plasmid conjugation

Nalidixic resistant (NalR) colonies of 644-PT8 were isolated from overnight culture on LB-agar with 20ug/ml Nal. Conjugation was performed on LB-agar by co-streaking donor strain 180-PT54 and recipient spontaneous NalR of 644-PT8. Co-streaked growth harvested in Phosphate-buffered saline then plated onto LB-agar with 20ug/ml Nal and 10ug/ml Cm. Resistant colonies purified by streaking onto fresh plates of 20ug/ml Nal and 10ug/ml Cm.

4.2.9 Acid Resistance Assays

Acid resistance assays were performed as described previously [147]. Briefly, cells were cultured overnight in either LBG (Luria-Bertani (LB) broth + 0.4 % glucose), LB buffered with 100 mM morpholinepropanesulfonic acid (MOPS pH8) or LB buffered with 100 mM morpholineethanesulfonic acid (MES pH 5.5). Overnight (22 h) stationary phase cultures were diluted 1:1000 into pre-warmed minimal E glucose (EG) media, pH 2.5. The glutamate and arginine dependent systems were tested by growing cells overnight in LBG and diluting cultures into EG pH 2.5 supplemented with either 1.5 mM glutamate or 0.6 mM arginine, respectively. The glucose repressed system was tested by growing cells overnight in LB-MES pH 5.5 followed by dilution in EG pH 2.5. Overnight cultures grown in either LB-MOPS (pH 8) or LBG followed by dilution in un-supplemented EG were used as acid-sensitive controls for the glucose-repressed and glutamate or arginine-dependent AR systems, respectively. Viable cells were enumerated at t = 0 and t = 4 h and used to calculate % survival.

4.2.10 Fitness assays

Fitness of 180-PT54 relative to 644-PT8 was calculated as described previously [148]. Viable-cell counts for each competing strain were determined at time zero (t =0) and again after 24 h of co-culturing by selective plating. Fitness was calculated using the formula:

Fitness index (f.i.) = LN (N_i (1)/ N_i (0)) / LN (N_j (1)/ N_j (0)),

Where N_i (0) and N_i (1) = initial and final colony counts of strain 180-PT54, respectively and

N_j (0) and N_j (1) = initial and final colony counts of strain 644-PT8, respectively[148].

4.2.11 Biolog Phenotyping Microarray

A single isolated Shiga toxin-containing *Escherichia coli* O157:H7 colony was grown on BUG+B agar overnight at 33°C. A sterile swab was used to transfer cells from the plate into inoculating fluid 0 (IF-0) to a turbidity of 43% T (transmittance) and addition IF-0 with dye was to a final cell density of 85% T. For phenotyping microarray (PM) plates 1 and 2 (Biolog, Inc., Hayward, CA), 100 ul/well was added. PM plates 3, 4, 6, 7 and 8 were supplemented with 20mM sodium succinate and 2 uM ferric citrate before 100 ul was added to each well [149]. All plates were incubated at 33°C for 48 hours using the Omnilog II Combo System (Biolog, Inc., Hayward, CA). The output data from the Omnilog was imported into the opm package in R for analysis [122].

4.3 Results

4.3.1 Short read sequencing Heatmap analysis

The clustering of gene families highlighted 8001 gene families in the pangenome of the 20 strains and of these 4082 were found in all strains so were regarded as the core genome. The remaining 3919 gene families were only found in subsets of the 20 strains so were considered to be members of the accessory genome. This was a very large proportion of accessory vs core genome and indicated that finding discriminatory gene differences between different PTs would be difficult. The heatmap (fig. 4.2) comparison showed that the differences between PTs were both plentiful and ambiguous. This is likely to be due to the poor assemblies that STEC strains produce when using short read sequencing. High phage content of STEC strains makes assembly very difficult. The repetitive nature of phage content makes reads difficult to place within the genome and results in a large number of short contigs that cannot be resolved. Long read sequencing was employed to explore on an examplar pair of strains from an outbreak that were isogenic and epidemiologically linked in an attempt to resolve this issue.

4.3.2 Short read sequencing analysis demonstrates the outbreak strains are closely related and were not endemic

All 145 cultures were received at the reference laboratory from cases linked to both the August and October clusters and were typed by phage typing and MLVA [40, 135]. The outbreak MLVA profile for all the outbreak associated strains isolated in August and those from cases in October was ?-8-12-4-5-2-8-3. The profile showed a high degree of variation at VNTR locus #3 but all isolates were an Single Locus Variant (SLV) of each other. Isolates that have the same MLVA profile or a SLV of that profile are regarded as microbiologically linked [40]. MVLA analysis had revealed that isolates from both the PT54 and PT8 outbreaks clustered together and that, despite their different PTs, the two occurrences of STEC O157 at the food outlet were likely linked so an investigation into their degree of relatedness was undertaken.

Analysis of Illumina sequences indicated that all four isolates from the August PT8 outbreak had identical core genome sequences. There were three-SNP differences in the core genome between the PT8 isolates from August and the PT54 isolates from October and two of these SNPs were unique to the August PT8 cluster. The maximum distance between isolates within the October PT54 cluster was 4 SNPs, including acquisition of a maximum of 2 SNPs from a common haplotype. BEAST analysis [150] of STEC WGS data predicts ~2.5 SNPs per year. Therefore it is likely that their last common ancestor was ~1 year before these public health incidents [6]. The unique SNPs observed in the isolates belonging to each cluster indicated that the PT54 cluster did not directly evolve from the PT8 cluster rather that they share a very recent common ancestor. Furthermore, it was evident that the PT8 and PT54 strains were closely related to each other but genetically distinct from strains of STEC O157 circulating in the local population (fig. 4.3). Strains held in the Public Health England (PHE) STEC O157 WGS database that clustered most closely with the outbreak strains were associated with foreign travel to Egypt and Israel (fig. 4.4). Although the precise source was never identified by the large investigation that followed the outbreaks, the strain is likely to have been imported in contaminated food with the larger outbreak exacerbated by an infected food handler in the restaurant.

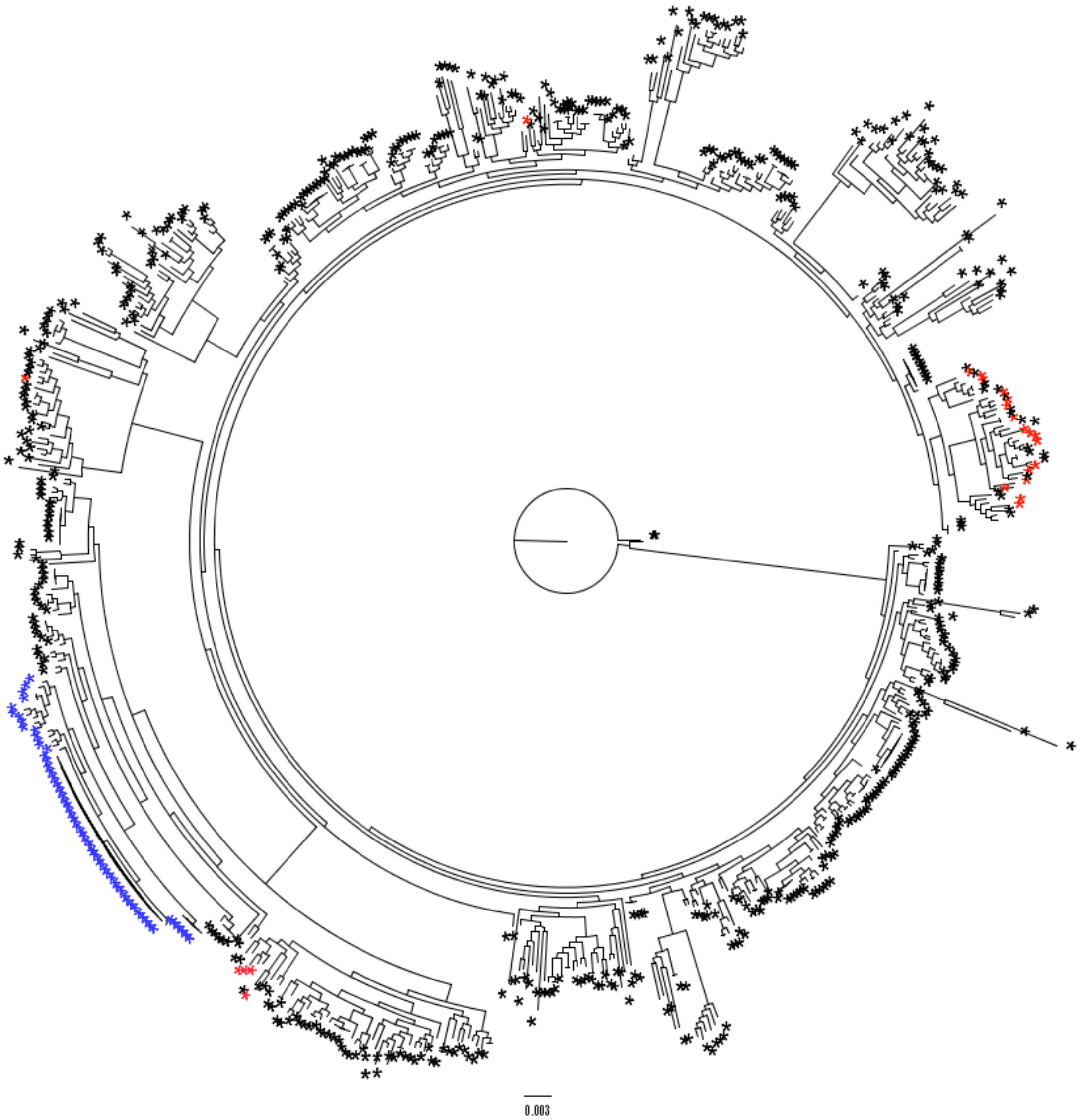


Figure 4.3. Maximum likelihood phylogenetic tree of STEC O157 strain selection of various PTs in PHE database associated with domestically acquired infection and travel related cases. SNPs called on core genome via a mapping technique against the reference strain Sakai. Outbreak strains shown in blue and local background strains in the region where the outbreak occurred are shown in red.

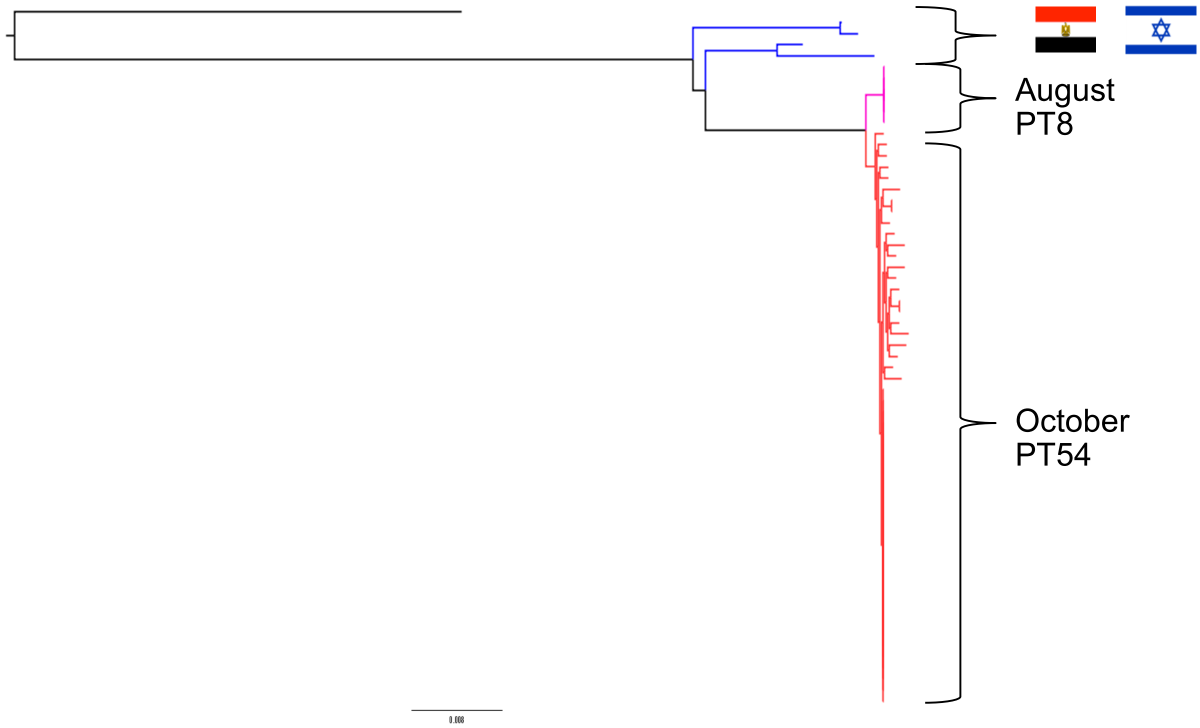


Figure 4.4. Maximum likelihood phylogenetic tree of outbreak strains and most closely related strains in the Colindale database that are associated with foreign travel to Egypt and Israel. The blue branches represent strains that are associated with travel to Egypt and Israel, the pink branches represent the original PT8 outbreak and the red branches represent the later PT54 outbreak.

4.3.3 PT8 and PT54 finished genomes

The PacBio sequencing assembled the genome of the PT54 isolate 180 into one contig and the genome of PT8 isolate 644 into two contigs. Difficulties with the assembly of isolate 644 were a result of the reorganized synteny of the genome compared to the closely related PT54 isolate.

From comparison of the MinION assembly and the OpGen map, it was clear that the assembly was distrupted because of a 200kbp inverted repeat in the genome that constituted the second smaller contig in the assembly. The contig was inserted twice into the larger contig to line up the NCol sites found in the OpGen map. The results of this can be seen in figure 4.5. It was found that the combination of the MinION sequencing assembly and the OpGen map enabled us to gain a single contig for the strain. The OpGen map of isolate 644 revealed that it was 5.8mb in length and the addition of the 200kbp repeat resolved the final assembly to that length.

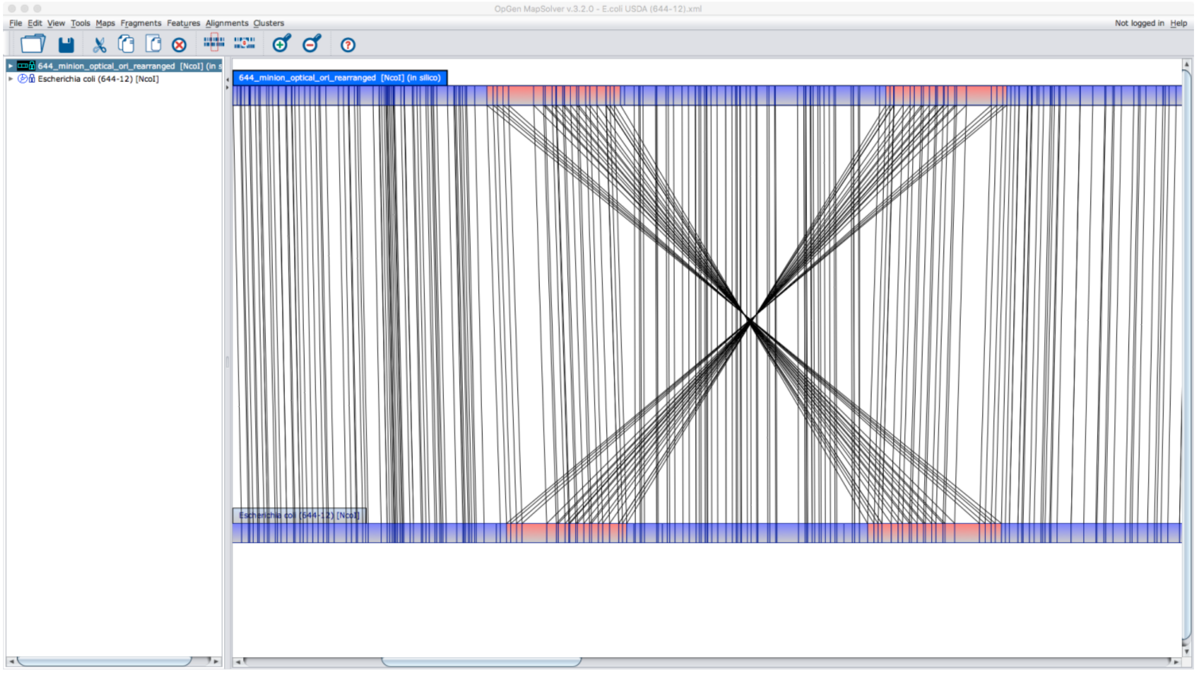

OpGen MapSolver v.3.2.0 - E.coli USDA (644-12).xml
File Edit View Tools Maps Fragments Features Alignments Clusters
Not logged in Help
644_minion_optical_ori_rearranged [NcoI] (in s
Escherichia coli (644-12) [NcoI]
644_minion_optical_ori_rearranged [NcoI] (in silico)
Escherichia coli (644-12) [NcoI]

Figure 4.5. OpGen map alignment with MinION assembly of the PT8 strain with smaller contig inserted twice. The NCol sites are joined up by single black lines clearly showing two regions with the same NCol sites that can each be mapped to two regions. All NCol sites have lined up nicely indicating a correct assembly.

4.3.4 Prophage profiling and chromosomal differences between specific PT8 and PT54 isolates based on long read sequencing methods

14 prophage regions were shared between the representative isolates of the PT8 and PT54 clusters (fig. 4.6). In addition to the 14 shared prophage regions, 180-PT54 had gained one small prophage region of 24,874bp located at 2281433-2306307bp and 644-PT8 had acquired one small prophage region of 20,818bp located at 4773172-4793990bp. In addition, the genome of 644-PT8 had three repeated prophages within the 200kbp inverted repeat (fig. 4.6). Two shared prophage

regions, designated P7 and P8 were similar prophages that showed variation between the two representative isolates, indicative of recombination that had contributed to the inverted repeat (figs. 4.6 and 4.7).

The unique gene differences within the chromosome of each strain are listed in appendix table 7.2. They are also represented in a schematic diagram (fig. 4.6). 644-PT8 had 82 unique genes on the chromosome, 79 of which were prophage associated. 180-PT54 had 30 unique genes, 27 of which were prophage associated. The non-prophage associated unique genes were all hypothetical proteins apart from an *ilvB* operon leader peptide found only in 180-PT54.

Whole genome alignment using NUCMER identified 29 SNPs between the two isolates (644-PT8 and 180-PT54) in fully aligned regions. The SNP locations and base changes are detailed in table 4.1. This was higher than the 3 SNPs identified between the 'core' genomes based on the short read sequencing but 26 of these SNPs were found in the P7 and P8 shared prophage regions; providing further evidence that these regions are recombination hotspots.

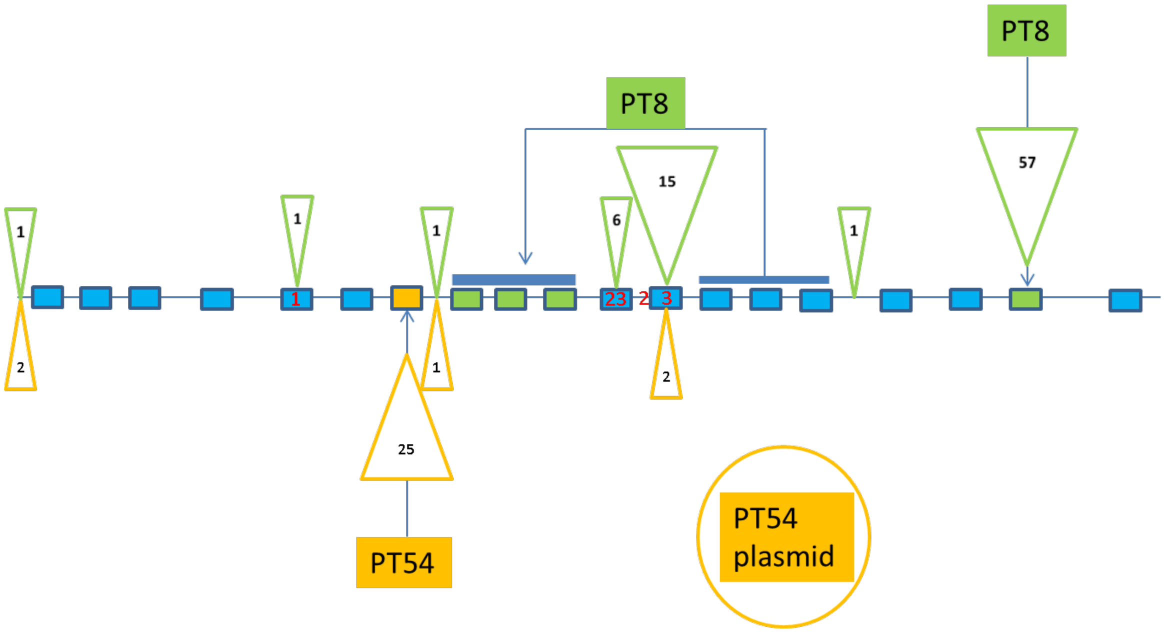


Figure 4.6. Schematic diagram representing accessory genome variation between PT8 and PT54 strains. Blue squares represent shared prophage regions, the orange square represents a unique PT54 prophage region and green squares represent unique PT8 prophage regions. Orange triangles represent location and number of unique genes for the PT54 strain and green triangles represent location and number of unique genes for the PT8 strain. SNPs are represented in red on the central line. Inverted repeat region underlined. PT54 strain unique plasmid represented by orange circle.

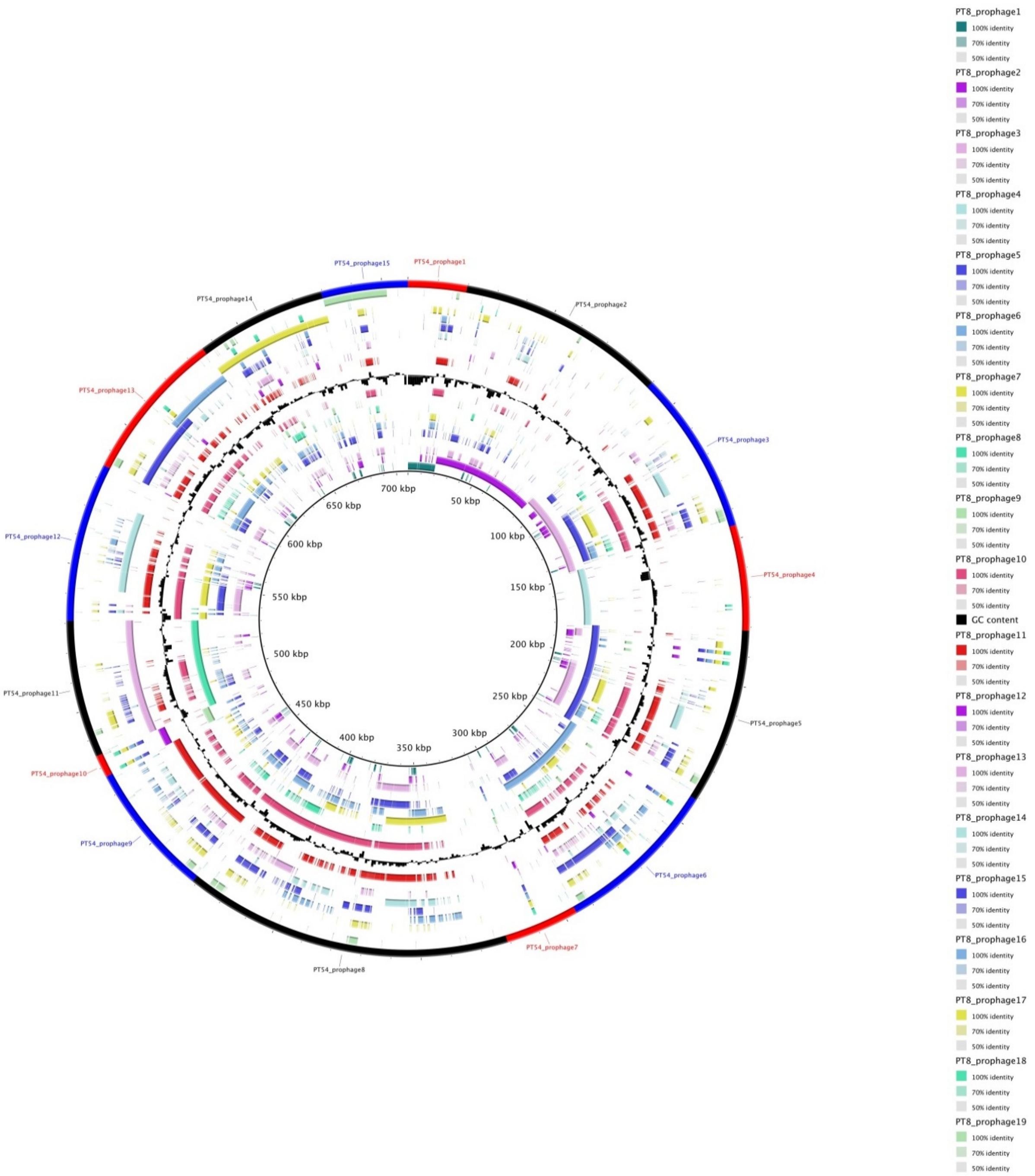


Figure 4.7. BRIG plot showing genomic similarity between PT54 strain prophage regions and PT8 strain prophage regions. Central ring is Multifasta of PT54 prophage regions(labelled on the outside) and each concentric ring represents 1 prophage region of the PT8 strain (labelled along right-hand side). The darker the colour the greater the level of genomic similarity in that part of the region. GC content is represented in black in the 11th ring.

Table 4.1. Table listing the position and base change of all the SNPs found between the PT8 strain and PT54 strain in a whole genome alignment using the program NUCMER. The SNPs identified by the previous phylogenetic analysis using Illumina data are highlighted in bold, the third SNP identified by the phylogenetic analysis is within a repeat region so was excluded as ambiguous alignment by the program NUCMER.

PT54 position	PT54 base	PT8 base	PT8 position
1975308	G	C	1974495
2681706	C	A	3282472
2681715	T	C	3282463
2681722	T	C	3282456
2681730	T	G	3282448
2681757	T	G	3282421
2681766	G	T	3282412
2681775	A	G	3282403
2681784	C	T	3282394
2681788	G	A	3282390
2681796	T	C	3282382
2681823	G	A	3282355
2681833	G	A	3282345
2681835	C	T	3282343
2681844	A	G	3282334
2681847	G	A	3282331
2681865	G	A	3282313
2681973	C	T	3282205
2681976	C	G	3282202
2681977	T	C	3282201

2681979	T	A	3282199
2681981	A	C	3282197
2681982	C	A	3282196
2681985	C	A	3282193
2833323	**A**	**C**	**3027880**
2894348	**A**	**G**	**3088905**
3048043	C	A	3242600
3048059	C	A	3242616
3048069	G	A	3242626

4.3.5 Plasmid acquisition by strain 180 (PT54)

Both strains harboured pO157, the O157 virulence plasmid present in nearly all strains of STEC O157 [151]. However, isolate 180-PT54 acquired an additional 220 genes introduced on an IncHI2 plasmid [152] not present in 644-PT8. While both 180-PT54 and 644-PT8 exhibited tellurite and tetracycline resistance, 180-PT54 was also resistant to chloramphenicol and streptomycin and this correlated with resistance genes located on the IncHI2 plasmid. The IncHI2 plasmid was predicted to encode at least six membrane proteins, a drug efflux pump, other resistance mechanisms including additional tellurite resistance and protection from exposure to heavy metal ions (mercury). It also encoded at least two DNA methylases (fig. 4.8). Relatedness depicted in a BRIG plot (fig. 4.8) demonstrates the high similarity with other IncHI2 plasmids that have been detected in clinical isolates worldwide [153-157].

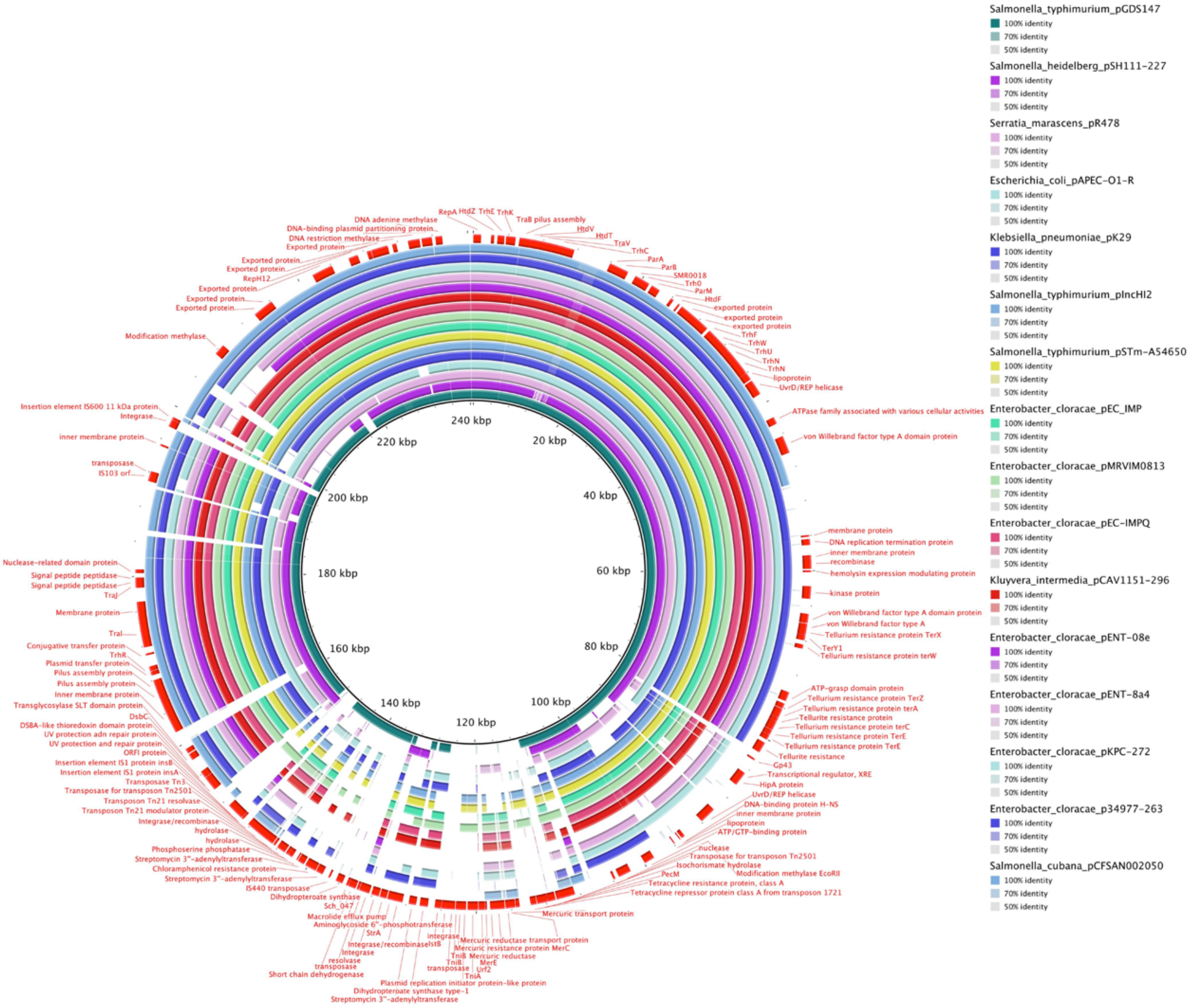

240 kbp
20 kbp
40 kbp
60 kbp
80 kbp
100 kbp
120 kbp
140 kbp
160 kbp
180 kbp
200 kbp
220 kbp
RepA
HtdZ
TrhE
TrhK
TraB pilus assembly
HtdV
HtdT
TraV
TrhC
ParA
ParB
SMR0018
Trh0
ParM
HtdF
exported protein
exported protein
exported protein
TrhF
TrhW
TrhU
TrhN
TrhN
lipoprotein
UvrD/REP helicase
ATPase family associated with various cellular activities
von Willebrand factor type A domain protein
membrane protein
DNA replication termination protein
inner membrane protein
recombinase
hemolysin expression modulating protein
kinase protein
von Willebrand factor type A domain protein
von Willebrand factor type A
Tellurium resistance protein TerX
TerY1
Tellurium resistance protein terW
ATP-grasp domain protein
Tellurium resistance protein TerZ
Tellurium resistance protein terA
Tellurite resistance protein
Tellurium resistance protein terC
Tellurium resistance protein TerE
Tellurium resistance protein TerE
Tellurite resistance
Gp43
Transcriptional regulator, XRE
HipA protein
UvrD/REP helicase
DNA-binding protein H-NS
inner membrane protein
lipoprotein
ATP/GTP-binding protein
nuclease
Transposase for transposon Tn2501
Isochorismate hydrolase
Modification methylase EcoRII
PecM
Tetracycline resistance protein, class A
Tetracycline repressor protein class A from transposon 1721
Mercuric transport protein
Mercuric reductase transport protein
Mercuric resistance protein MerC
Mercuric reductase
MerE
Urf2
TniA
TniB
TniB
transposase
integrase
IstB
Plasmid replication initiator protein-like protein
Dihydropteroate synthase type-1
Streptomycin 3"-adenylyltransferase
Integrase/recombinase
Integrase
resolvase
transposase
Short chain dehydrogenase
StrA
Aminoglycoside 6"-phosphotransferase
Macrolide efflux pump
Sch_047
Dihydropteroate synthase
IS440 transposase
Streptomycin 3"-adenylyltransferase
Chloramphenicol resistance protein
Streptomycin 3"-adenylyltransferase
Phosphoserine phosphatase
hydrolase
hydrolase
Integrase/recombinase
Transposon Tn21 modulator protein
Transposon Tn21 resolvase
Transposase for transposon Tn2501
Transposase Tn3
Insertion element IS1 protein insA
Insertion element IS1 protein insB
ORFI protein
UV protection and repair protein
UV protection adn repair protein
DSBA-like thioredoxin domain protein
DsbC
Transglycosylase SLT domain protein
Inner membrane protein
Pilus assembly protein
Pilus assembly protein
Plasmid transfer protein
TrhR
Conjugative transfer protein
TraI
Membrane protein
TraJ
Signal peptide peptidase
Signal peptide peptidase
Nuclease-related domain protein
IS103 orf
transposase
inner membrane protein
Integrase
Insertion element IS600 11 kDa protein
Modification methylase
Exported protein
Exported protein
Exported protein
RepH12
Exported protein
Exported protein
Exported protein
DNA restriction methylase
DNA-binding plasmid partitioning protein
DNA adenine methylase
Salmonella_typhimurium_pGDS147
100% identity
70% identity
50% identity
Salmonella_heidelberg_pSH111-227
100% identity
70% identity
50% identity
Serratia_marascens_pR478
100% identity
70% identity
50% identity
Escherichia_coli_pAPEC-O1-R
100% identity
70% identity
50% identity
Klebsiella_pneumoniae_pK29
100% identity
70% identity
50% identity
Salmonella_typhimurium_pIncHI2
100% identity
70% identity
50% identity
Salmonella_typhimurium_pSTm-A54650
100% identity
70% identity
50% identity
Enterobacter_cloracae_pEC_IMP
100% identity
70% identity
50% identity
Enterobacter_cloracae_pMRVIM0813
100% identity
70% identity
50% identity
Enterobacter_cloracae_pEC-IMPQ
100% identity
70% identity
50% identity
Kluyvera_intermedia_pCAV1151-296
100% identity
70% identity
50% identity
Enterobacter_cloracae_pENT-08e
100% identity
70% identity
50% identity
Enterobacter_cloracae_pENT-8a4
100% identity
70% identity
50% identity
Enterobacter_cloracae_pKPC-272
100% identity
70% identity
50% identity
Enterobacter_cloracae_p34977-263
100% identity
70% identity
50% identity
Salmonella_cubana_pCFSAN002050
100% identity
70% identity
50% identity

Figure 4.8. BRIG plot of 240kbp IncHI2 plasmid found in PT54 strain as central reference showing genomic similarity between it and other IncHI2 plasmids found in Genbank from various species of bacteria. Annotations shown in red on outermost ring. The darker the colour the greater the level of genomic similarity between the PT54 IncHI2 plasmid and the plasmid found in a different organism. The plasmid each colour refers to is labelled vertically on the right hand side.

4.3.6 Phage type transition associated with plasmid acquisition

It was hypothesized that the IncHI2 plasmid may be responsible for the difference in PT observed between the two outbreak clusters in August and October. To test this, the IncHI2 plasmid was conjugated into 644-PT8 and the conjugant was then phage typed. Acquisition of the plasmid, as defined by inheritance of chloramphenicol resistance, converted 644-PT8 to PT54. Antibiotic resistance profiling demonstrated that the plasmid-associated resistance was present in all the PT54 isolates in this study indicating they all carried the IncHI2 plasmid associated with the PT transition. Analysis of the genes present on the plasmid (fig. 4.8) shows a number of determinants that could be associated with changes in phage resistance including tellurite resistance [158]. While both 644-PT8 and 180-PT54 have chromosomally-encoded Tellurite resistance, specifically *terW*, only 180-PT54 have *terY* and *terX* as these are present on the IncHI2 plasmid. Methylase modification genes encoded on the plasmid are also potential candidates to confer resistance to specific bacteriophages [73].

4.3.7 Increased fitness of the PT54 strain associated with the larger second outbreak

Acquisition of IncHI2 plasmids commonly confers phage, antibiotic and heavy metal resistance thus increasing bacterial fitness under certain environmental conditions [159, 160]. Conversely an increased metabolic burden imposed by the 200Kbp inversion during DNA replication in strain 644-PT8 is likely to reduce fitness. To assess any differences in fitness, strains 644-PT8 and 180-PT54 were competed by co-culturing in LB-broth at 37°C and 25°C. Under both conditions 180-PT54 significantly outcompeted 644-PT8 (Fitness index = 1.28 and 1.23, respectively). Fitness was therefore independent of culture temperature. Subsequently Biolog phenotypic microarrays were performed to determine the nature of the observed fitness increase. Comparable growth was observed for both strains for the majority of carbon, phosphorus and sulphur sources tested (data not shown) however growth of 180-PT54 was enhanced for multiple nitrogen sources, including certain amino acids and dipeptides (fig. 4.9). This is in agreement with the observed increased fitness of 180-PT54 when cultured in LB-broth in which amino acids/short peptides are the primary carbon and nitrogen sources. In addition, growth of 180-PT54 was better than 644-PT8 when ammonia was the sole nitrogen source (fig. 4.9). Of the multiple di- and tri-peptide nitrogen sources on which 180-PT54 grew better than 644-PT8 many contained arginine or glutamate as part of their composition. *E. coli* possesses three acid resistance systems (AR) of which two are dependent on arginine and glutamate, respectively [161]. It was therefore tested if AR was altered in 180-PT54 by the increased metabolism of arginine and glutamate relative to 644-PT8. For each AR system (glucose repressed, arginine and glutamate dependent) strain 644-PT8 was significantly more resistant to acid shock when either pre-adapted in LB pH 5.5 or supplied with exogenous Arg or Glu (table 4.2). Without pre-adaptation however strain 180-PT54 was more acid resistant than 644-PT8 (p = 0.035).

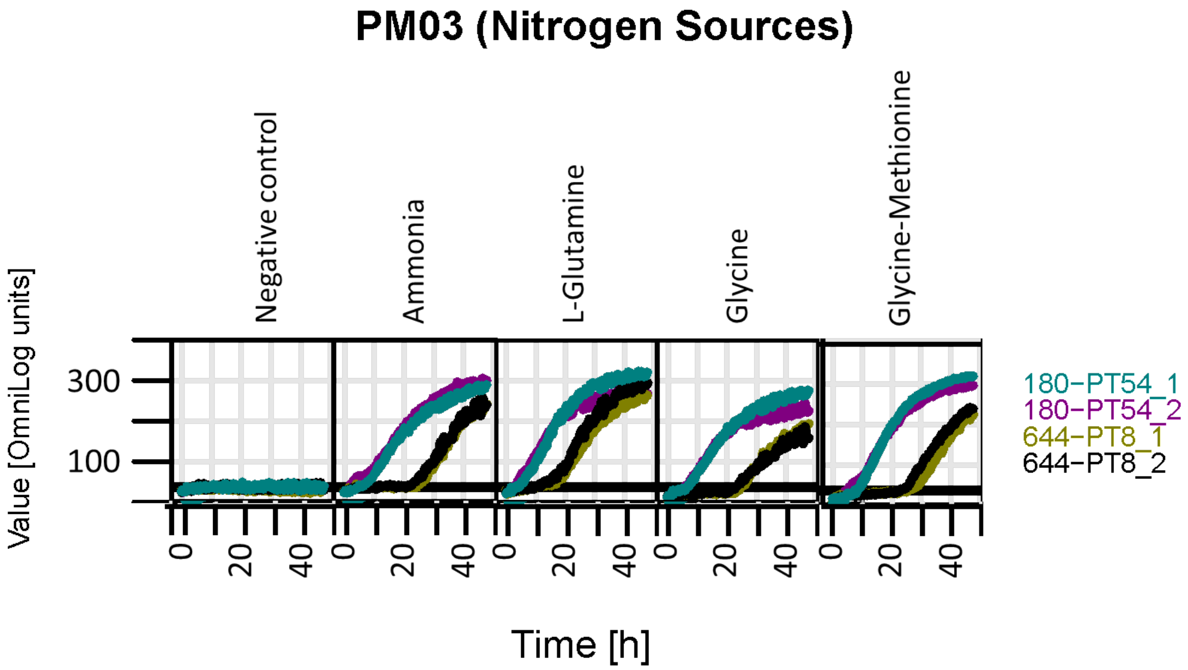


Figure 4.9. XY Plots of respiration curves from Shiga toxin-containing *Escherichia coli* strains 180 and 644 grown in four nitrogen sources from Phenotype MicroArray plate PM03 and the negative control. Each strain was run in duplicate on different day and indicated by the different colored curves. The growth time in hours is on

the x-axis and reduction of the reporter dye was quantified as OmniLog units are represented on the y-axis

Table 4.2. Table describing the results of the acid resistance assays performed on strains 644-PT8 and 180-PT54 to assess their biological fitness.

		% Survival[b] (Standard Deviation)	
Adaptation medium[a]	Challenge Medium (pH2.5)	PT54	PT8
LB pH8	EG	0.81 (0.45)	0.19 (0.19)
LB pH 5.5	EG	4.10 (2.74)	38.79 (3.39)
LBG	EG	14.77 (8.51)	7.23 (5.08)
	EG+Glu	21.72 (4.12)	52.70 (8.06)
	EG+Arg	37.28 (8.79)	70.11 (10.37)

[a] Strains were adapted by overnight culturing in either LB pH5.5 or LBG before diluting 1:1000 in EG pH2.5 or supplemented EG.[b] The percentage survival was determined after 4 h of acid challenge. The mean % survival of six replicates (n = 6) is shown for EG challenge and three replicates (n = 3) for EG supplemented with either Glutamate (1.5 mM) or Arginine (0.6 mM). The standard deviation for the replicates is represented below the % survival in brackets. Without pre-adaptation strain 180-PT54 was more acid resistant than 644-PT8 (*p* = 0.035).

4.4 Discussion

The range of STEC O157 strains of various PT that were selected for Illumina sequencing and genome comparison, through a heatmap to look for gene association with PT, indicated that long read sequencing would facilitate this study. To define gene association with specific PTs good evidence of specific genes/alleles being the key difference or similarity between strains that are different or the same PT respectively was required. The accessory genome for STEC O157 is very large so differences between even closely related strains are multiple, even the exemplar isogenic strains that were long read sequenced had 82 and 30 genes unique to each of them. For this reason long read sequencing was used to look at mobile elements such as prophages or plasmids and other powerful molecular techniques such as TraDIS (chapter 5) to discriminate and provide a selection for genes that are involved in PT were employed.

Phylogeny techniques based on 'core genome' sequence analyses have made a significant contribution and provide added insight in to epidemiological investigations and also provide an assessment of evolutionary relationships between strains. Analysis of STEC O157 WGS data indicated that strains with less than 5 SNP differences in the core genome are likely to be epidemiologically linked [6] and this can be determined from short read sequencing. Strains from two temporally related outbreaks of STEC O157 from the same restaurant were examined. Initially, MLVA and phage typing results were contradictory as MLVA indicated the outbreaks were caused by the same strain although the phage types were distinct. The relatedness of the PT8 and PT54 strains was confirmed by short read sequencing which defined 3 SNP differences in the core genome between the two groups of strains indicating that they share a recent common ancestor (within 1 year) (fig. 4.4).

The application of PacBio and MinION sequencing, as well as OpGen mapping, produced single contig assemblies of two isolates associated with the two outbreaks at the restaurant, one belonging to PT8 and one belonging to PT54. These assemblies clearly showed the high prophage carriage in these isolates, which is typical of STEC O157 with at least 20% of the genome made up of highly paralogous phage genes (fig. 4.7). These data illustrate why short read sequencing cannot be used to accurately define the location and composition of related prophage regions. Analysis of the genomes from the long read sequencing showed that there had been a shift in prophage composition between the two outbreak groups. There was gain and loss of prophage while two of the shared prophage regions had undergone recent recombination. Furthermore, isolate 644-PT8 had a repeat of three of the shared prophage regions in a 200kbp inverted region. There was significant, apparent functional redundancy across the different prophages. This is in agreement with earlier research from Hayashi and colleagues that showed a diverse bacteriophage complement produced from a single strain including recombination between prophage loci [57]. This analysis, based on insights from the two different sequencing technologies, provided the first evidence of prophage microevolution in STEC O157 between two closely related outbreaks.

Isolate 180-PT54 assembled into three contigs; the chromosome, the F-like pO157 and an IncHI2 plasmid. The IncHI2 plasmid was large (240Kbp) and predicted to encode about 220 genes. This plasmid was not present in 644-PT8. IncHI2 plasmids are commonly associated with the spread of extended spectrum β-lactam resistance (ESBL) genes, heavy metal resistance and phage resistance (Whelan, Colleran et al. 1995, Fang, Li et al. 2016). Several antibiotic and environmental resistance genes were identified, potentially conferring resistance to chloramphenicol, streptomycin, tellurite, tetracycline and certain heavy metal ions encoded on the incHI2 plasmid acquired by 180-PT54 (fig. 4.8). The resistance loci facilitated conjugation experiments of the plasmid into 644-PT8, leading to the recipient strain phage typing as PT54. Conversion of PT8 to PT54 is caused by the acquisition of resistance to the group 3 typing phages (TP4, TP5 and TP14) [162]. There is a previous report that an IncHI2 plasmid can confer bacteriophage resistance [158] therefore adding to the possible survival advantage conferred on strains that acquire this plasmid. A BLAST search revealed that highly similar IncHI2 plasmids have been described in several different organisms from around the world (fig 4.8), including Taiwan, China, Kenya, Malawi and the USA. Identification of incHI2 plasmids in *E. coli* however is rare [160, 163]. The acquisition of this plasmid is therefore likely to increase the survival capacity of the strain under certain stressful environmental conditions.

In addition to antibiotic and phage resistance, 180-PT54 was shown to be significantly fitter than 644-PT8 under a defined set of growth conditions. Although this could not be solely attributed to the acquisition of the incHI2 plasmid, the genetic differences identified in 180-PT54 that include plasmid acquisition resulted in a fundamental alteration in central nitrogen metabolism in strain 180-PT54 that enhanced growth over 644-PT8. In accordance with the trade-off between self-preservation and nutritional competency (SPANC) however the increased growth of 180-PT54 has resulted in a decrease in acid resistance, at least under priming conditions [164]. The public health investigation of the outbreaks included sampling of employees who worked at the restaurant. Two members of staff were shown to be colonized with the PT54 strain during the outbreak and an analysis identified a significant association with one of these employees and the infection risk, although it is not known if this individual could have actually accounted for the second outbreak. Based on this and the altered genotype and phenotype of the PT54 strain, it was proposed that it may be more adapted for human colonization than the original PT8 strain and this capacity may be linked to the much higher number of cases associated with the second outbreak, including multiple examples of human-to-human, in contact, transmission. STEC O157 strains are usually associated with ruminant hosts and presumably human colonization could lead to adaptive changes that promote survival such as plasmid acquisition and prophage variation. Sequencing of human cases in the UK has identified a subset of imported strains that are acquired during travel abroad or brought in by colonised foreign visitors [6]. These strains are significantly more likely to contain plasmids encoding antibiotic resistance and there is a concern that acquisition of such elements may adapt strains to then persist in the human population, which would be a serious public health concern. Therefore PTs that are determined by plasmid acquisition, as this chapter has demonstrated for the PT54 phenotype, could also have certain survival advantages over other strains. This would mean that PT could be seen in certain environments more commonly than others or could be more associated with human-to-human transmission.

4.5 Conclusions

This chapter has highlighted the issues of identifying genetic determinants of PT from short read sequencing data. The long read sequencing data from two closely related outbreak strains was highly useful and allowed for the recognition of the IncHI2 plasmid association with the PT54 phenotype. The use of long read WGS data clearly demonstrated the dynamic nature of the accessory genome in STEC O157 and the potential impact of horizontal gene transfer over a short time frame. The long read sequencing enabled us to identify a plasmid that confers resistance to antibiotics and bacteriophages as well as other environmental stressors. It is clear that both short and long read sequencing can greatly facilitate outbreak investigations into foodborne gastrointestinal diseases caused by pathogens with dynamic genomes but that long read sequencing is preferential when investigating phage type conversion.

Chapter 5: The use of TraDIS to elucidate genes involved in typing phage susceptibility

5.1 Background

Analysis of the genome sequencing data in the previous chapters revealed that there was a multitude of prophage differences in closely related strains that were obscured by Illumina short read sequencing, and highlighted issues with elucidating the genetic determinants of PT from incomplete assemblies. To overcome this, methods involving phenotypic selection to elucidate genetic mechanisms were researched and Transposon Directed Insertion-site Sequencing (TraDIS) was highlighted as a successful method previously used to investigate phage susceptibility [89]. TraDIS is a genome-wide mutagenesis method that when used with phenotype screening can identify loci that contribute to the selection phenotype.

TraDIS involves the production of a high-density transposon insertion library by random introduction of a transposon into the genome. The random insertion enables a high number of mutants to be generated containing multiple insertions in every non-essential gene and therefore represents a comprehensive mutant library for selection purposes. The mutant library is then used for selection under a specific condition/pressure. Genes and other loci that promote or reduce survival under those conditions are then identified by transposon-directed sequencing [87]. Selection can be identified based on adjusted read numbers for a particular insertion.

The aim of this chapter was to investigate mechanisms of phage resistance and susceptibility by generating a TraDIS library of the STEC O157 strain 9000. By exposing this high-density transposon library to a typing phage to which it is susceptible to, I hypothesized that the recovered population should contain gene insertions that alter the level of resistance to the phage.

5.2 Methods

5.2.1 Library Strain selection

A PT32 Stx 2a/2c knockout of strain 9000 termed strain 1465 (chapter 3) was used to produce the library. A derivative of strain 9000 was chosen because of the high number of previous studies done on the strain providing information about its characteristics *in vivo* [165].

5.2.2 Transposon used

A transposon kit from Epicentre was used; the EZ-Tn5 <KAN-2> Tnp Transposome kit. It is a stable complex of the EZ-Tn5 transposase enzyme and the EZ-Tn5 <KAN-2> Transposon that contains the Tn903 kanamycin resistance gene. Once inside the *E. coli* cells the transposase is activated by Mg2+ and randomly inserts into the host genome. Tn5 transposons are preferable to mariner-based transposons because they do not require a target sequence for insertion so a potentially greater insertion density can be reached [86].

5.2.3 Producing Competent cells of Library strain 1465

Two single colonies of strain 1465 were inoculated into a 20ml volume of LB broth each and left to grow overnight at 37^{0}C. Each 20ml overnight culture was added to 300ml of LB broth and incubated at 37^{0}C until they reached OD 0.5. The cultures were then centrifuged at 3000 rcf for 10 mins at 4^{0}C in 6 x 50ml aliquots. The supernatant was decanted and the cells were suspended in 25ms of ice-cold 10% glycerol. The cells were then centrifuged again at 3000 rcf for 10mins at 4^{0}C. For the second time, the supernatant was decanted, the cells were resuspended in 12.5ml of ice-cold 10% glycerol and centrifuged at 3000 rcf for 10mins at 4^{0}C.

Finally, the supernatant was removed and the cells were resuspended in 6.25ml of ice-cold 10% glycerol and centrifuged at 3000 rcf for 10mins at 4^{0}C.

Subsequently the supernatant was removed and the cells were resuspended in 0.5ml of ice-cold 10% glycerol and transferred to 1.5ml microfuge tubes. The cells were then centrifuged at 9000 rpm for 10mins. Again the supernatant was removed, the cells were resuspended in 50ul of ice-cold 10% glycerol and kept on ice before electroporation.

5.2.4 Electroporation of transposome into competent cells

0.5ul of the EZ-Tn5 transposome was added to each of the 6 x 50ul of competent cells on ice. The transposome and competent cells mixture was then transferred into 0.1cm electroporation cuvettes and electroporated in a Bio-Rad gene pulser. Cells were immediately recovered in 1ml of SOC broth after electroporation. This was repeated for each of the 6 aliquots of competent cells. The recovered cells in SOC broth were then incubated for 2 hours at 37^{0}C on a shaking incubator.

5.2.5 Recovery of transformed cells

Each of the 6 electroporated cultures were spread on a 250ml LB plates containing kanamycin (4mg/L) and incubated overnight at 37^{0}C. Colonies of transformed cells were harvested in the morning after plate count and stored at -80^{0}C in a final concentration of 20% glycerol after the OD had been recorded for that batch. This method was repeated 30 times to produce 30 different batches and >1,000,000 mutant colonies had been harvested.

5.2.6 Library pooling

Batches were pooled according to optical density as shown in table 5.1.

Table 5.1 A table to detail the optical density, date cultured and amount pooled for the TraDIS library.

Batch(date)	O.D.	Pool(ml)
1(29/5/14)	2.49	0.25
2(6/6/14)	2.93	0.3
3(11/6/14)	0.434	1.25
4(25/6/14)	2.808	0.3
5(27/6/14)	2.832	0.3
6(2/7/14)	1.617	0.75
7(25/7/14)	0.362	1.25
8(30/7/14)	2.479	0.25
9(1/8/14)	0.465	1.25
10(27/8/14)	1.283	0.75
11(29/8/14)	0.937	0.5
12(3/9/14)	3.02	0.3
13(5/9/14)	2.334	0.2
14(10/9/14)	2.568	0.25
15(12/9/14)	1.779	0.2
16(17/9/14)	1.705	0.2
17(19/9/14)	2.892	0.3

18(24/9/14)	2.595	0.25
19(14/1/15)	0.835	0.5
20(15/1/15)	0.795	0.5
21(16/1/15)	1.755	0.2
22(4/2/15)	0.883	0.5
23(5/2/15)	0.618	1
24(6/2/15)	2.279	0.27
25(25/2/15)	2.631	0.26
26(26/2/15)	1.34	0.4
27(27/2/15)	2.293	0.2
28(4/3/15)	1.487	0.4
29(5/3/15)	1.155	0.4
30(6/3/15)	2.746	0.28
Total		13.76

5.2.7 TraDIS sequencing and library density

TraDIS specific sequencing was performed at the Wellcome Trust Sanger Institute (WTSI) to produce transposon-directed reads. To determine insertion sites of each transposon that had randomly inserted into the genome of each mutant, a special TraDIS library prep was used that uses specially designed TraDIS adapters and primers to increase the enrichment of genuine transposon-chromosome junctions by preventing hybridization of the reverse primer until the transposon-specific forward primer has generated a complementary strand [166]. This was to ensure that the first 10bp of every read was transposon sequence and the remaining sequence was downstream of where the transposon was inserted. These reads were then mapped using SMALT(WTSI) and insertion quantification was performed to a PacBio sequenced reference for strain 9000 to determine the location the transposon has inserted at. Strains that had transposon insertions in essential genes did not survive and therefore the sequence was only recovered for non-essential genes. Analysis of the sequencing data showed that the library had >110,000 unique transposon insertion sites throughout the genome with a density of one insertion every 50bp of the genome demonstrating that there were multiple unique insertions in every viable gene and functionally significant non-coding regions. The multiple sites in each locus together act as independent indicators (similar to 'biological replicates') [86].

5.2.8 Bacteriophage selections on library

The library was subjected to bacteriophage 13 of the typing phages which was fully lytic on a PT32 library, the phage type of the variant strain 9000. A pool of the library (see table 5.1) was cultured overnight with 10^6 cells added to 10ml of LB broth. The following morning 100ul of the library was inoculated into 10ml of LB broth and allowed to grow up to log phase. 100ul of the log phase culture was then inoculated into three x 10ml of LB broth. Ten ul and 200ul (multiplicity of infection ~10) of bacteriophage were inoculated into two of the broths and one used as a control without bacteriophage selection but cultured in parallel in LB for the same time. DNA extractions were taken at both three and five hour time points using the Promega Wizard extraction kit. DNA extractions for the controls and the selections were sent to the WTSI for TraDIS-specific sequencing (see table 5.2).

5.2.9 Bioinformatics analysis

Transposon directed sequencing reads were mapped, using the SMALT(WTSI) alignment algorithm, to a complete genome sequence (sequenced using PacBio technology) of strain 9000 (see table 5.2). A change in the number of reads that mapped to each gene between the control and the selections was measured by LogFC (log2 fold change) calculated from log counts per million. This was done using publicly available WTSI TraDIS analysis scripts (https://github.com/sanger-pathogens/Bio-Tradis) that automatically determines appropriate read mapping parameters from the length of the first read in the fastq file. Log2 fold was used as a measure of comparison of read number fold changes compared to the LB control ->2 and <-2 were used as a cut off for genes with different numbers of insertions[166]. P and Q values are given for each calculation as a measure of statistical significance in terms of the false positive (P) and false discovery (Q) rate [167]. Only those that had a Q value of <0.01 and P value of <0.05 were counted (see figs 5.1 and 5.2). Results from both time points were averaged out for both concentrations to give a consensus list of genes involved in phage susceptibility and resistance, present at both time points and concentrations.

Table 5.2 Sequencing results of TraDIS specific sequencing performed on bacteriophage selections and controls, detailing the number of sequenced reads and the perentage of those reads that mapped to the reference strain 9000. The total unique insertion sites indicate the number of transposon mutants present when DNA extraction took place.

Sample name	Total Reads	% Mapped	Total Unique Insertion Sites
3_control	4381188	89.425229	182227
3_P13_10ul	4749002	86.7993153	164218
3_P13_200ul	4556057	90.7622474	138883
5_control	5011685	90.7141662	186851
5_P13_10ul	5194424	92.4274558	127199
5_P13_200ul	4140930	95.7753279	82909

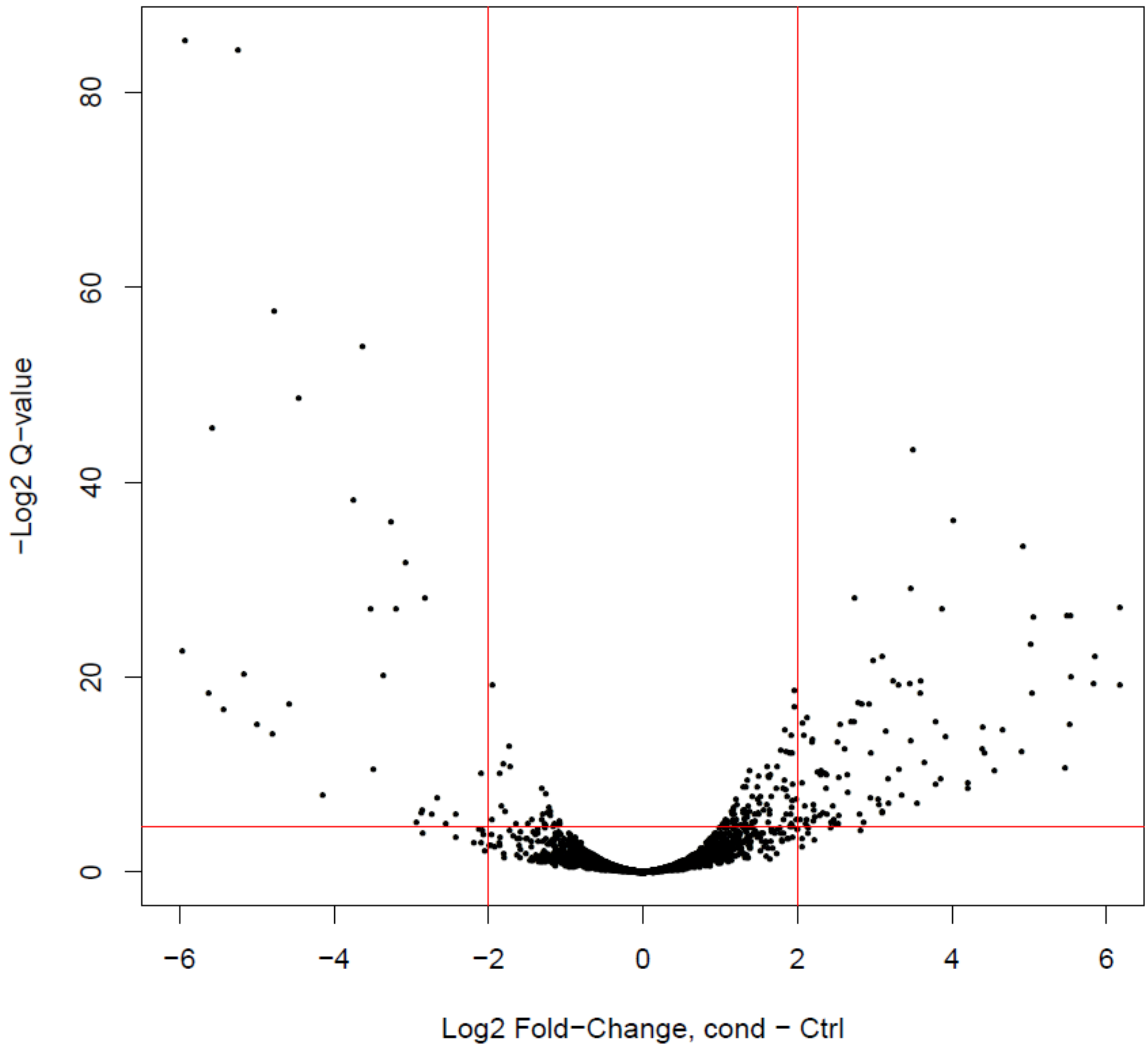


Figure 5.1 Volcano plot showing change in prevalence of mutants from the control to the addition of 10ul phage selections. As shown by the relationship of Log2 Fold change in selection condition compared to the control (x-axis) with Q-value(y-axis) indicating false discovery rate. Red lines show the cutoff criteria of 10% false discovery rate (horizontal) and a log2 FC of 2 (vertical).

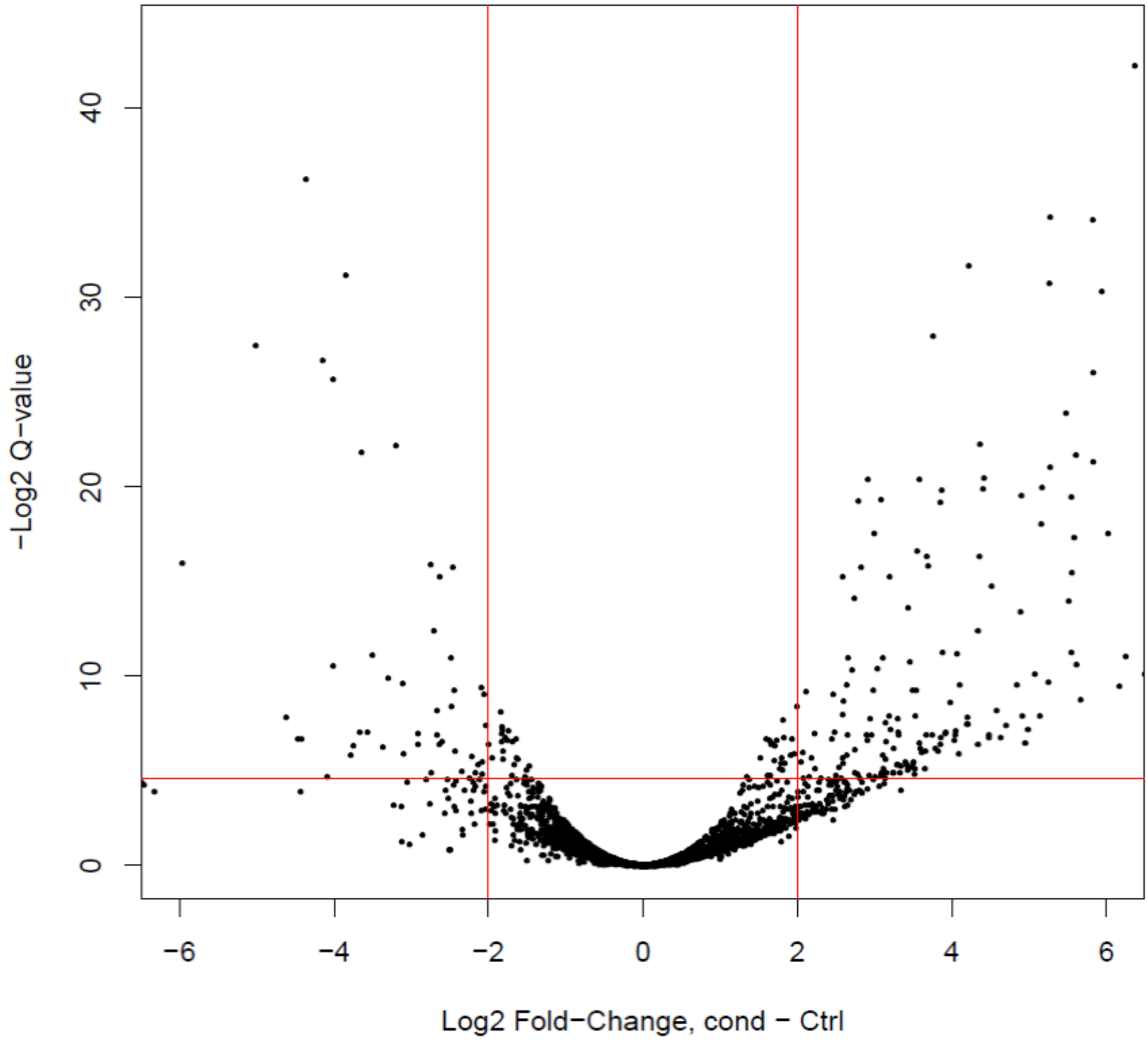


Figure 5.2 Volcano plot showing change in prevalence of mutants from the control to the addition of 200ul phage selections. As shown by the relationship of Log2 Fold change in selection condition compared to the control (x-axis) with Q-value (y-axis) indicating false discovery rate. Red lines show the cutoff criteria of 10% false discovery rate (horizontal) and a log2 FC of 2 (vertical).

5.2.10 COG annotation

COG is the clustering of orthologous genes in gene families associated with known functions [168] and is a database that can be downloaded from the internet. This database can then be queried using blastp [119] for COG hits that can assign your query protein to a group of orthologous proteins. The blast cutoffs for assignment were >90% identity and coverage.

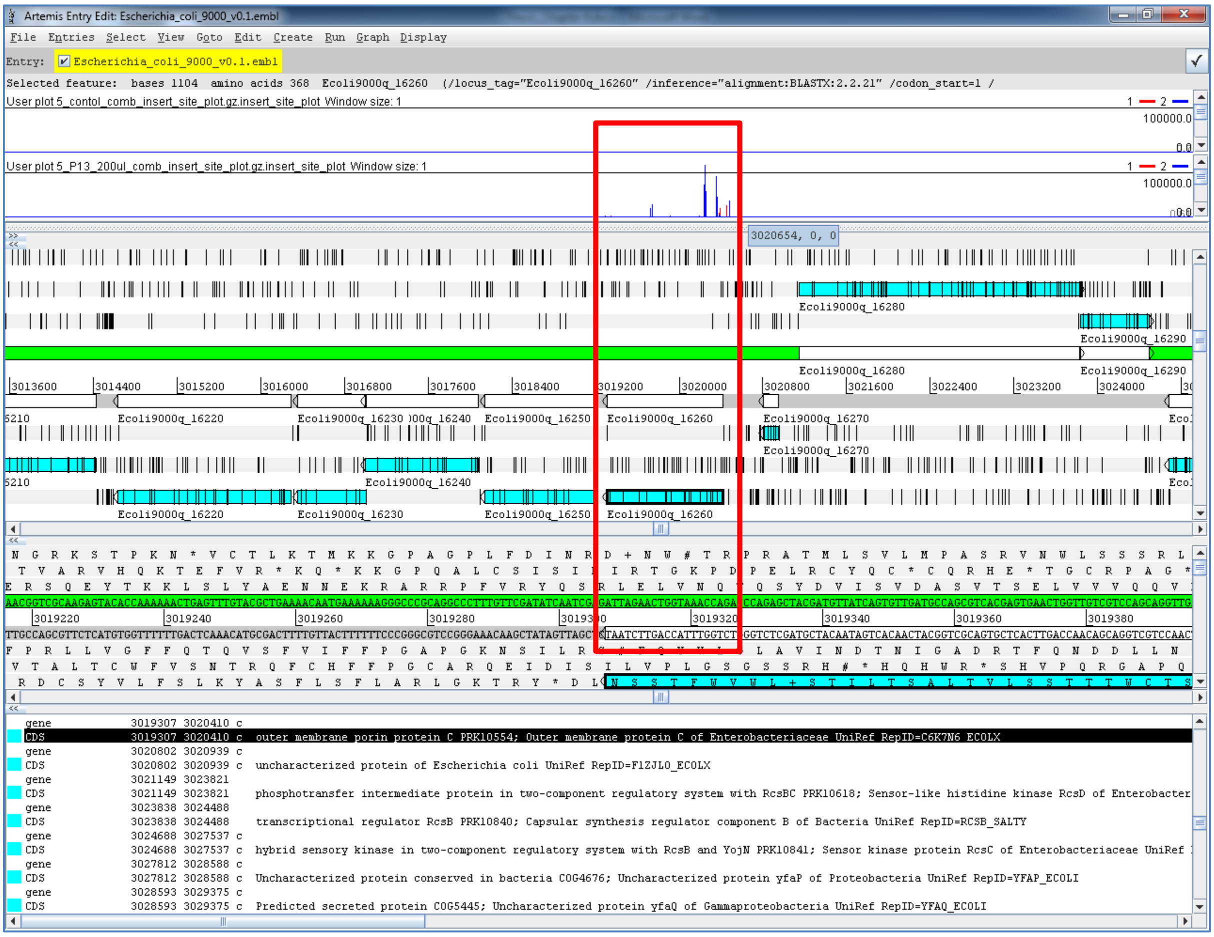
5.3 Results

5.3.1 Genes involved in phage sensitivity

The genes shown in appendix table 7.3 with LogFC >2 contain insertions that led to the mutants growing at a faster rate than the control (greater than 2 fold) in the presence of phage. This means these clones were not being infected or killed as readily by the selecting phage. These mutants may be important for phage adherence/infection and involved a total of 114 genes. With a nine fold increase in insertions, OmpC is the most important gene for T4 phage infection that is universally found in *E. coli* (fig 5.3).

The 114 genes were found in almost all regions of the chromosome (fig 5.4) and were involved in multiple pathways and aspects of the host’s functions (fig 5.5) as indicated by their COG assignment. The dispersed location of these genes and the involvement of multiple functional pathways in bacteriophage infection indicate that bacteriophages have evolved on multiple occasions to use increasing numbers of the host’s functional pathways for efficient lysis.

The involvement of the Sap operon (fig 5.6) in bacteriophage infection is of particular interest because it is known to be associated with antimicrobial resistance [169]. This TraDIS selection data indicates that the Sap operon while protecting the strain from antibiotics is also exposing the strain to more successful bacteriophage infection.



```
Selected feature:  bases 1104  amino acids 368  Ecoli9000q_16260  (/locus_tag="Ecoli9000q_16260" /inference="alignment:BLASTX:2.2.21" /codon_start=1 /
```

```
gene    3019307 3020410 c
CDS     3019307 3020410 c  outer membrane porin protein C PRK10554; Outer membrane protein C of Enterobacteriaceae UniRef RepID=C6K7N6 ECOLX
gene    3020802 3020939 c
CDS     3020802 3020939 c  uncharacterized protein of Escherichia coli UniRef RepID=F1ZJL0_ECOLX
gene    3021149 3023821
CDS     3021149 3023821    phosphotransfer intermediate protein in two-component regulatory system with RcsBC PRK10618; Sensor-like histidine kinase RcsD of Enterobacter
gene    3023838 3024488
CDS     3023838 3024488    transcriptional regulator RcsB PRK10840; Capsular synthesis regulator component B of Bacteria UniRef RepID=RCSB_SALTY
gene    3024688 3027537 c
CDS     3024688 3027537 c  hybrid sensory kinase in two-component regulatory system with RcsB and YojN PRK10841; Sensor kinase protein RcsC of Enterobacteriaceae UniRef
gene    3027812 3028588 c
CDS     3027812 3028588 c  Uncharacterized protein conserved in bacteria COG4676; Uncharacterized protein yfaP of Proteobacteria UniRef RepID=YFAP_ECOLI
gene    3028593 3029375 c
CDS     3028593 3029375 c  Predicted secreted protein COG5445; Uncharacterized protein yfaQ of Gammaproteobacteria UniRef RepID=YFAQ_ECOLI
```

Figure 5.3 Artemis insertion plot showing the number of reads corresponding to transposon insertions present in each gene for the infection assay. As shown in this screen shot there are nearly 100,000 reads for transposon insertions in OmpC after the 200ul phage selection (bottom insertion plot) indicating this clone has grown well during bacteriophage challenge compared to the control (upper insertion plot) which has far less insertions present. The number of insertions is represented by lines in the insertion plots as highlighted by the red box.

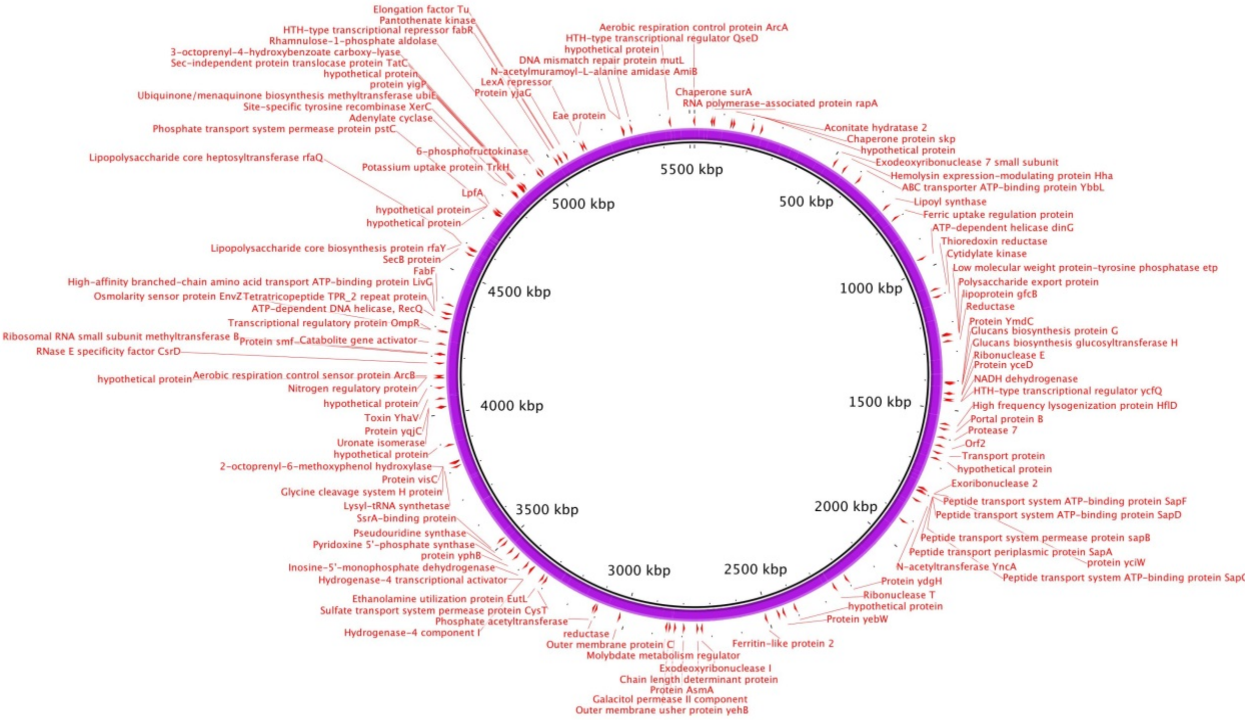

Elongation factor Tu
Pantothenate kinase
HTH-type transcriptional repressor fabR
Rhamnulose-1-phosphate aldolase
3-octoprenyl-4-hydroxybenzoate carboxy-lyase
Sec-independent protein translocase protein TatC
hypothetical protein
protein yigP
Ubiquinone/menaquinone biosynthesis methyltransferase ubiE
Site-specific tyrosine recombinase XerC
Adenylate cyclase
Phosphate transport system permease protein pstC
6-phosphofructokinase
Lipopolysaccharide core heptosyltransferase rfaQ
Potassium uptake protein TrkH
LpfA
hypothetical protein
hypothetical protein
Lipopolysaccharide core biosynthesis protein rfaY
SecB protein
FabF
High-affinity branched-chain amino acid transport ATP-binding protein LivG
Osmolarity sensor protein EnvZ
Tetratricopeptide TPR_2 repeat protein
ATP-dependent DNA helicase, RecQ
Transcriptional regulatory protein OmpR
Ribosomal RNA small subunit methyltransferase B
Protein smf
Catabolite gene activator
RNase E specificity factor CsrD
hypothetical protein
Aerobic respiration control sensor protein ArcB
Nitrogen regulatory protein
hypothetical protein
Toxin YhaV
Protein yqjC
Uronate isomerase
hypothetical protein
2-octoprenyl-6-methoxyphenol hydroxylase
Protein visC
Glycine cleavage system H protein
Lysyl-tRNA synthetase
SsrA-binding protein
Pseudouridine synthase
Pyridoxine 5'-phosphate synthase
protein yphB
Inosine-5'-monophosphate dehydrogenase
Hydrogenase-4 transcriptional activator
Ethanolamine utilization protein EutL
Sulfate transport system permease protein CysT
Phosphate acetyltransferase
Hydrogenase-4 component I
Aerobic respiration control protein ArcA
HTH-type transcriptional regulator QseD
hypothetical protein
DNA mismatch repair protein mutL
N-acetylmuramoyl-L-alanine amidase AmiB
LexA repressor
Protein yjaG
Eae protein
Chaperone surA
RNA polymerase-associated protein rapA
Aconitate hydratase 2
Chaperone protein skp
hypothetical protein
Exodeoxyribonuclease 7 small subunit
Hemolysin expression-modulating protein Hha
ABC transporter ATP-binding protein YbbL
Lipoyl synthase
Ferric uptake regulation protein
ATP-dependent helicase dinG
Thioredoxin reductase
Cytidylate kinase
Low molecular weight protein-tyrosine phosphatase etp
Polysaccharide export protein
lipoprotein gfcB
Reductase
Protein YmdC
Glucans biosynthesis protein G
Glucans biosynthesis glucosyltransferase H
Ribonuclease E
Protein yceD
NADH dehydrogenase
HTH-type transcriptional regulator ycfQ
High frequency lysogenization protein HflD
Portal protein B
Protease 7
Orf2
Transport protein
hypothetical protein
Exoribonuclease 2
Peptide transport system ATP-binding protein SapF
Peptide transport system ATP-binding protein SapD
Peptide transport system permease protein sapB
Peptide transport periplasmic protein SapA
protein yciW
N-acetyltransferase YncA
Peptide transport system ATP-binding protein SapC
Protein ydgH
Ribonuclease T
hypothetical protein
Protein yebW
Ferritin-like protein 2
reductase
Outer membrane protein C
Molybdate metabolism regulator
Exodeoxyribonuclease I
Chain length determinant protein
Protein AsmA
Galacitol permease II component
Outer membrane usher protein yehB
5500 kbp
500 kbp
1000 kbp
1500 kbp
2000 kbp
2500 kbp
3000 kbp
3500 kbp
4000 kbp
4500 kbp
5000 kbp

Figure 5.4 BRIG plot showing the location of 114 genes involved in bacteriophage infection on the chromosome of strain 9000

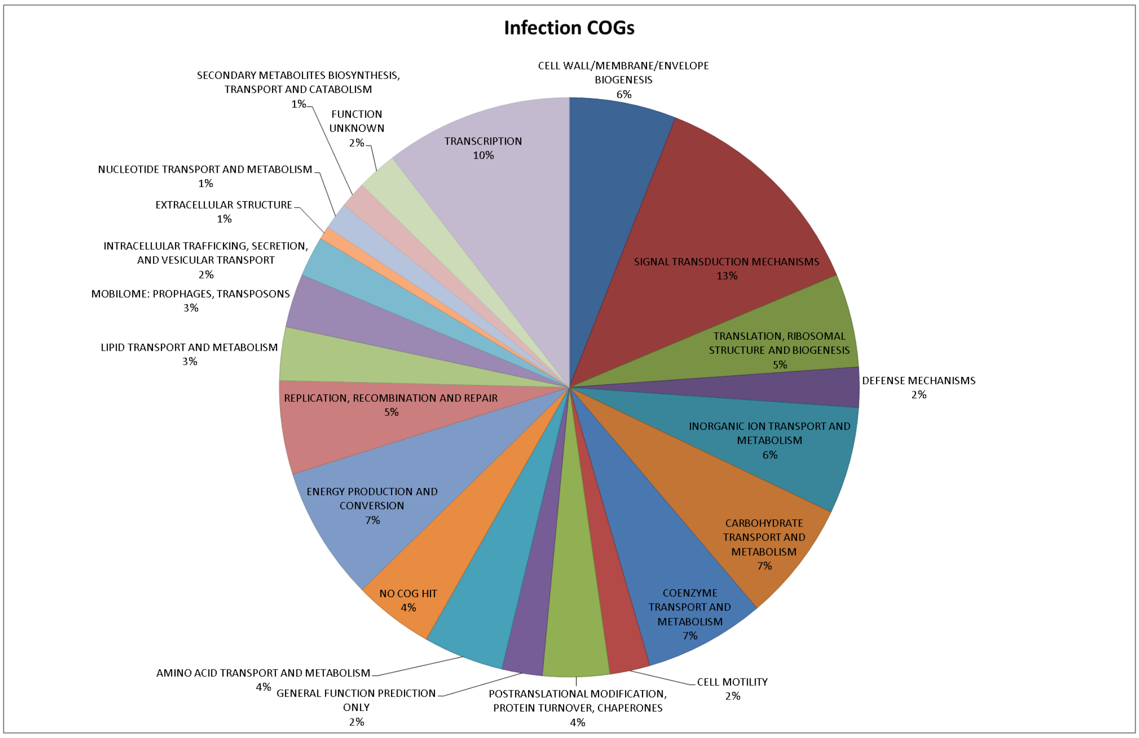


Figure 5.5 Pie chart displaying the percentage of different COGs represented in the genes shown by the TraDIS selections to be involved in bacteriophage infection

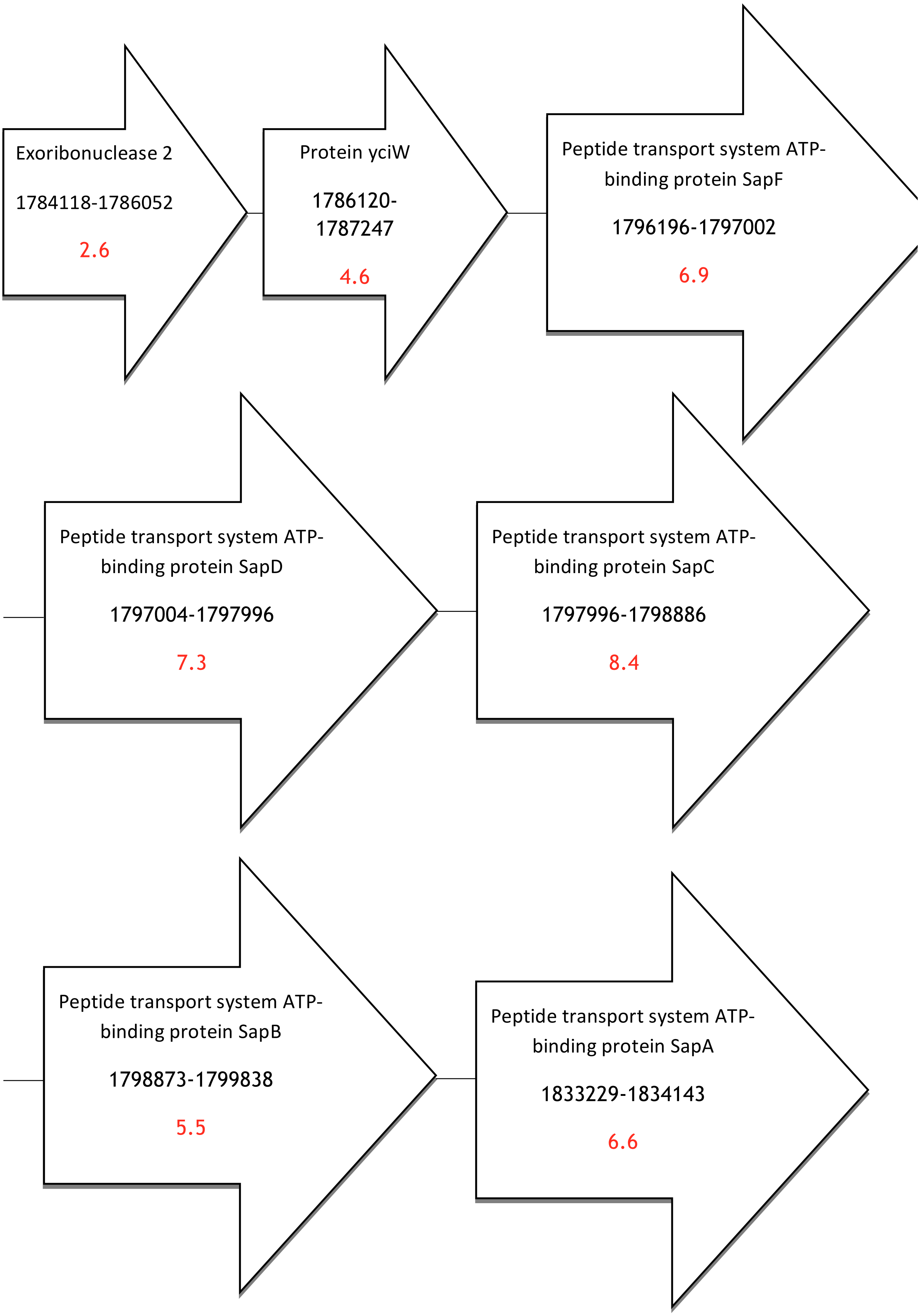

Exoribonuclease 2
1784118-1786052
2.6
Protein yciW
1786120-1787247
4.6
Peptide transport system ATP-binding protein SapF
1796196-1797002
6.9
Peptide transport system ATP-binding protein SapD
1797004-1797996
7.3
Peptide transport system ATP-binding protein SapC
1797996-1798886
8.4
Peptide transport system ATP-binding protein SapB
1798873-1799838
5.5
Peptide transport system ATP-binding protein SapA
1833229-1834143
6.6

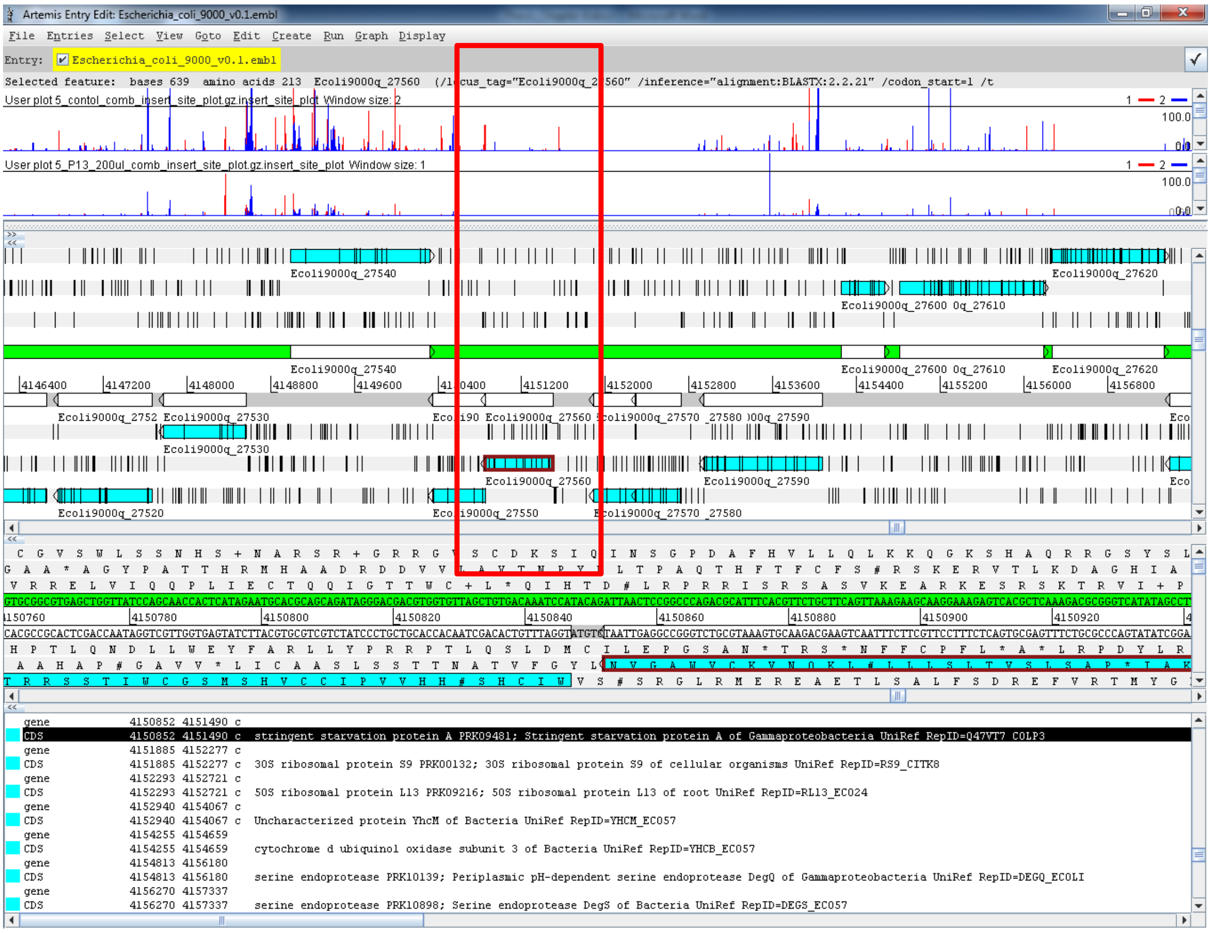
Figure 5.6 Schematic diagram of the Sap operon encoding a peptide transport system ATP-binding protein running from 1784118-1834143, this operon has previously been implicated in antibiotic resistance [169] and the whole operon has been implicated by this TraDIS study for successful bacteriophage infection. The arrows represent the gene coding regions with annotation, chromosomal location and log fold change in red (rounded to one decimal place).

5.3.2 Genes involved in phage resistance

The genes shown in appendix table 7.4 with LogFC >-2 contain insertions that result in the mutants growing at a lower rate (greater than 2 fold less) than the control in the presence of phage. This means these clones were lysed more frequently than other clones and so were infected or killed more readily by the selecting phage. These mutants may therefore be important for phage resistance. Forty-four candidate genes were detected and this small number was not surprising as the WT for the library is susceptible. The stringent starvation protein has the largest decrease in insertions compared to the control (see fig 5.7), and could be involved in the prevention of phage production by halting all transcription. The majority of the genes involved in resistance were associated with lipopolysaccharide (LPS) synthesis.

The 44 genes were found in multiple regions of the chromosome (fig 5.8) but were clustered in some regions indicating operon involvement in resistance. The majority of the genes were assigned to the cell wall/membrane/envelope biogenesis COG group (fig 5.9) indicating that the cell surface was the most important method of defending the cell against bacteriophage attack. The external surface of a bacterial cell is the first point of contact for an infecting bacteriophage through the outer membrane or LPS. The LPS may be an initial adsorption target but LPS O chains might also occlude OMPs. If a cell is able to occlude receptors found on these regions by the presence of LPS O chains or some other mechanism, then this could increase their resistance. Conversely, an insertion that inhibits O-antigen production may result in the mutant becoming more susceptible because the phage can gain direct access to main irreversible adsorption sites on OMPs.

The operon displayed in fig 5.10 is involved in carbohydrate transport and metabolism as well as cell wall/membrane/envelope biogenesis and all of the displayed genes were implicated in bacteriophage resistance by the TraDIS selection data. The operon also contains the O-antigen polymerase Wzy and the O-antigen ligase is also showing a log fold decrease on another region of the chromosome which indicated that the O-antigen itself is blocking phage adsorption in some manner.



Selected feature: bases 639 amino acids 213 Ecoli9000q_27560 (/locus_tag="Ecoli9000q_27560" /inference="alignment:BLASTX:2.2.21" /codon_start=1 /t

```
gene   4150852 4151490 c
CDS    4150852 4151490 c  stringent starvation protein A PRK09481; Stringent starvation protein A of Gammaproteobacteria UniRef RepID=Q47VT7 COLP3
gene   4151885 4152277 c
CDS    4151885 4152277 c  30S ribosomal protein S9 PRK00132; 30S ribosomal protein S9 of cellular organisms UniRef RepID=RS9_CITK8
gene   4152293 4152721 c
CDS    4152293 4152721 c  50S ribosomal protein L13 PRK09216; 50S ribosomal protein L13 of root UniRef RepID=RL13_EC024
gene   4152940 4154067 c
CDS    4152940 4154067 c  Uncharacterized protein YhcM of Bacteria UniRef RepID=YHCM_EC057
gene   4154255 4154659
CDS    4154255 4154659    cytochrome d ubiquinol oxidase subunit 3 of Bacteria UniRef RepID=YHCB_EC057
gene   4154813 4156180
CDS    4154813 4156180    serine endoprotease PRK10139; Periplasmic pH-dependent serine endoprotease DegQ of Gammaproteobacteria UniRef RepID=DEGQ_ECOLI
gene   4156270 4157337
CDS    4156270 4157337    serine endoprotease PRK10898; Serine endoprotease DegS of Bacteria UniRef RepID=DEGS_EC057
```

Figure 5.7 Artemis insertion plot graph showing the number of reads corresponding to transposon insertions present in each gene for the resistance assay. As shown in this screen shot, there are close to 0 transposon insertions present in Stringent starvation protein A after the 200ul phage selection compared to the control that has nearly 100 transposon insertions present in Stringent starvation protein A without phage selection. This indicates that this clone has struggled to survive during bacteriophage selection compared to the control which has far more insertions present in this gene. The number of insertions is represented by lines in the insertion plots as highlighted by the red box.

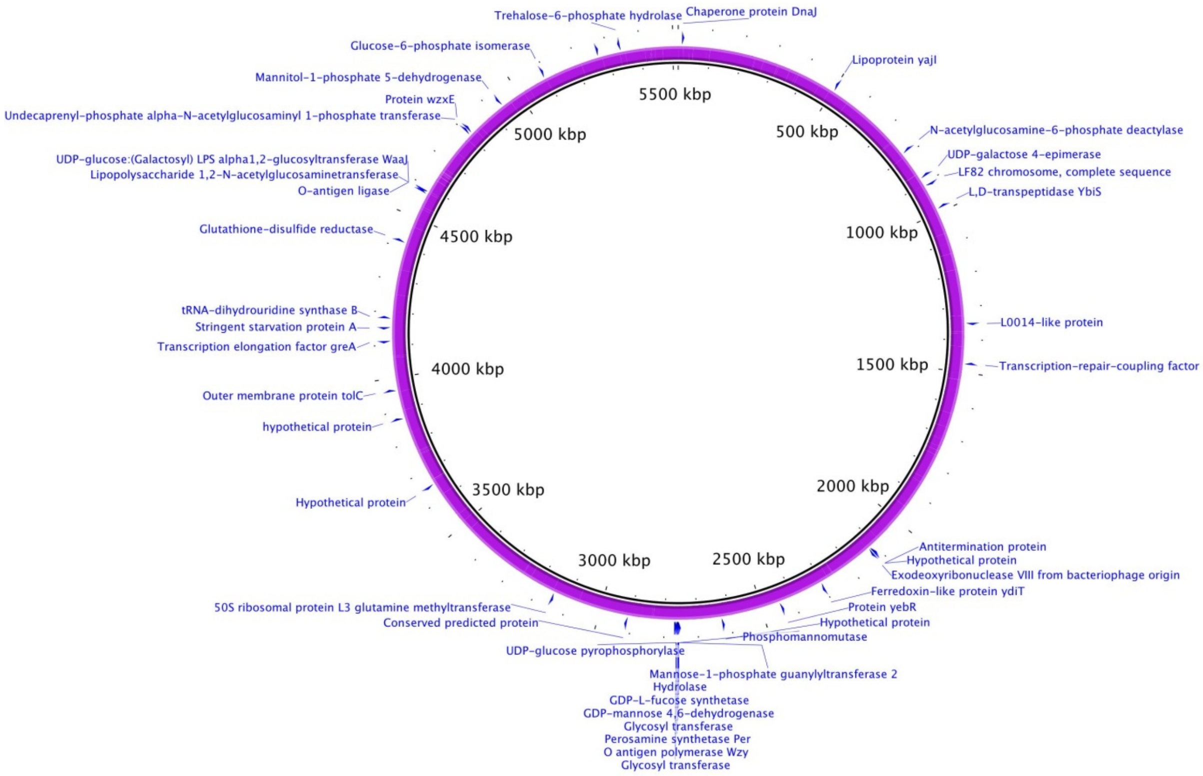

Trehalose−6−phosphate hydrolase
Chaperone protein DnaJ
Glucose−6−phosphate isomerase
Lipoprotein yajI
Mannitol−1−phosphate 5−dehydrogenase
Protein wzxE
Undecaprenyl−phosphate alpha−N−acetylglucosaminyl 1−phosphate transferase
5500 kbp
5000 kbp
500 kbp
N−acetylglucosamine−6−phosphate deactylase
UDP−galactose 4−epimerase
LF82 chromosome, complete sequence
L,D−transpeptidase YbiS
UDP−glucose:(Galactosyl) LPS alpha1,2−glucosyltransferase WaaJ
Lipopolysaccharide 1,2−N−acetylglucosaminetransferase
O−antigen ligase
Glutathione−disulfide reductase
4500 kbp
1000 kbp
tRNA−dihydrouridine synthase B
Stringent starvation protein A
Transcription elongation factor greA
L0014−like protein
4000 kbp
1500 kbp
Transcription−repair−coupling factor
Outer membrane protein tolC
hypothetical protein
Hypothetical protein
3500 kbp
2000 kbp
3000 kbp
2500 kbp
Antitermination protein
Hypothetical protein
Exodeoxyribonuclease VIII from bacteriophage origin
Ferredoxin−like protein ydiT
Protein yebR
Hypothetical protein
50S ribosomal protein L3 glutamine methyltransferase
Conserved predicted protein
Phosphomannomutase
UDP−glucose pyrophosphorylase
Mannose−1−phosphate guanylyltransferase 2
Hydrolase
GDP−L−fucose synthetase
GDP−mannose 4,6−dehydrogenase
Glycosyl transferase
Perosamine synthetase Per
O antigen polymerase Wzy
Glycosyl transferase

Figure 5.8 BRIG plot showing the location of the 44 TraDIS highlighted genes implicated in bacteriophage resistance on the strain 9000 chromosome.

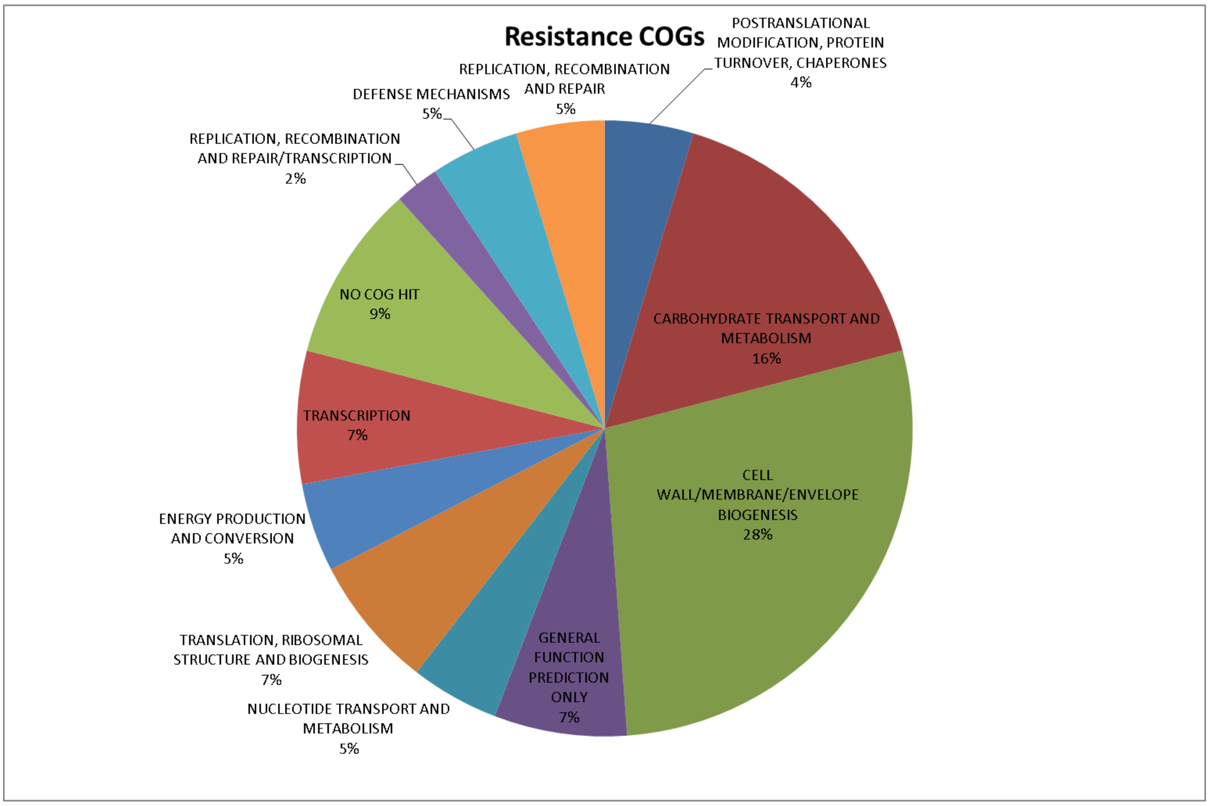


Figure 5.9 Pie chart displaying the percentage of different COGs represented in the genes shown by the TraDIS selections to be involved in bacteriophage resistance.

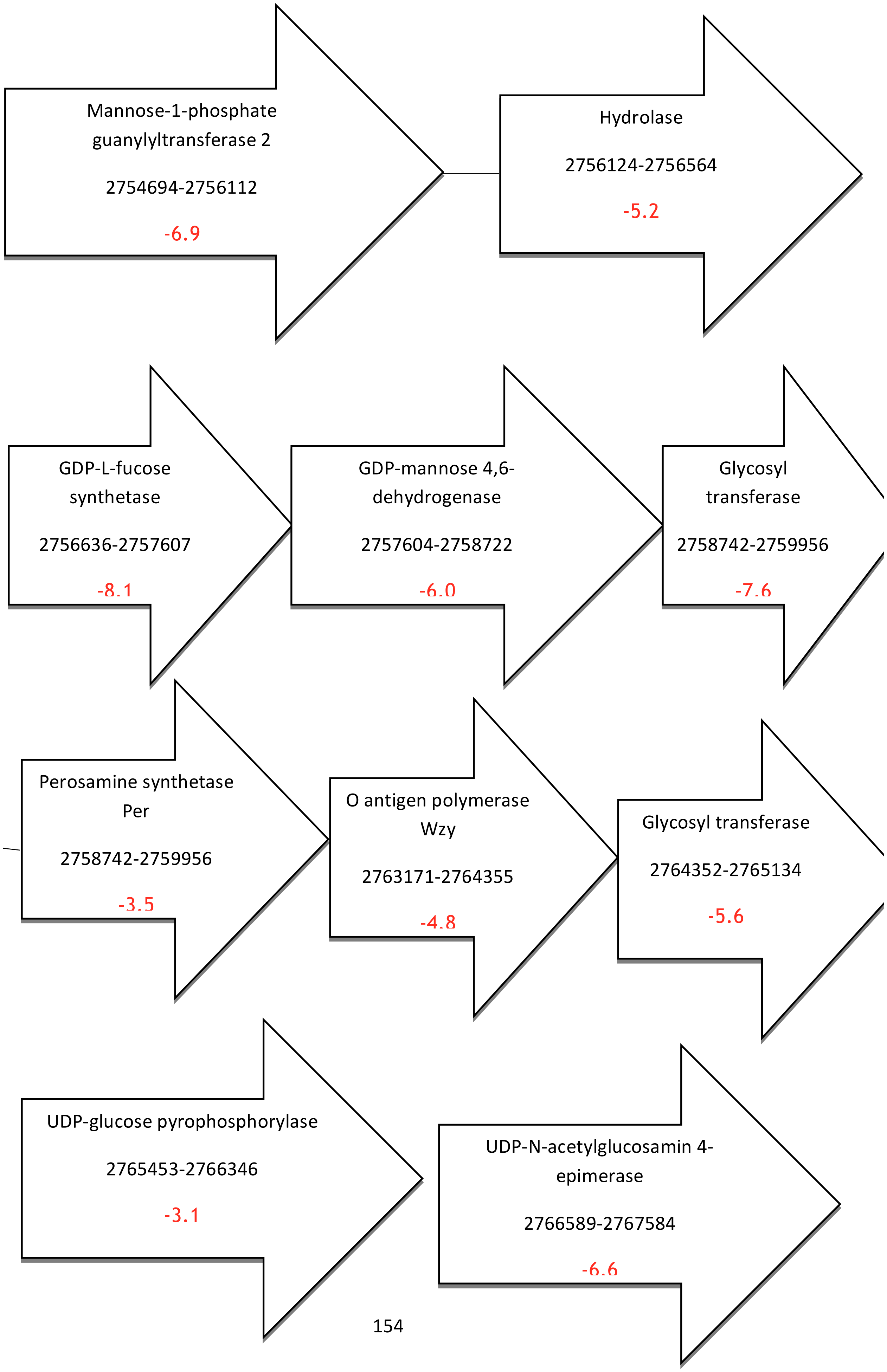

Mannose-1-phosphate guanylyltransferase 2
2754694-2756112
-6.9
Hydrolase
2756124-2756564
-5.2
GDP-L-fucose synthetase
2756636-2757607
-8.1
GDP-mannose 4,6-dehydrogenase
2757604-2758722
-6.0
Glycosyl transferase
2758742-2759956
-7.6
Perosamine synthetase Per
2758742-2759956
-3.5
O antigen polymerase Wzy
2763171-2764355
-4.8
Glycosyl transferase
2764352-2765134
-5.6
UDP-glucose pyrophosphorylase
2765453-2766346
-3.1
UDP-N-acetylglucosamin 4-epimerase
2766589-2767584
-6.6

Figure 5.10 An operon running from 2754694 – 2767584 that is involved in cell wall/membrane/envelope biogenesis and the O antigen was highlighted by the TraDIS selections as being involved in bacteriophage resistance. The arrows represent the gene coding regions with annotation, chromosomal location and log fold change in red.

5.4 Discussion

TraDIS is a genome-wide mutagenesis tool that has proved useful in looking for genes associated with PT conversion [89]. In this study TraDIS was used to investigate bacteriophage-host interactions between a single typing phage and one strain of STEC O157. In the TraDIS experiment reported in this chapter, survivors of bacteriophage selection changed from their wild type susceptible phenotype to a more resistant phenotype due to transposon inserts in specific genes that normally hinder phage infection. Using this method several genes involved in STEC O157 susceptibility and resistance to a T4 bacteriophage were successfully identified.

In section 5.3.1 the genes that had an increase in insertions compared to the control are listed in appendix table 7.3. These are genes that are associated with phage infection and are mainly outer membrane targets/receptors. The most influential gene, based on the log fold change in number of insertions, is OmpC. This is an established receptor for T4 in *E. coli* [170] and so without it mutants are able to flourish under phage infection compared to the remainder of the population. OmpC is a universal outer membrane protein in *E. coli* so T4 bacteriophage that use it as a receptor generally have a broad host range across *E. coli*; this interaction is reversible but initial infection and cell adherence is possible if the target bacterial cell has OmpC [171]. Closely following *ompC* mutants in terms of increased resistance are the transcriptional regulators OmpR and EnvZ, which are known to control *ompC* [172, 173]. The next most significant gene encodes a hypothetical protein that may have a role in translation, ribosomal structure and biogenesis by COG assignment. The involvement of this gene in phage infection requires further investigation. The data implicated the whole Sap operon (fig 5.6) as inserts in most genes of this operon were increased in the phage-challenged pool. The involvement of this operon in both antibiotic resistance and bacteriophage infection also requires further investigation. Also seen in the list of genes involved in infection and efficient lysis were genes involved in transcription and translation. It makes sense that successful phage production would benefit from biasing the transcriptional and translational apparatus of the host cell to phage products. It has yet to be established what role the other genes play in phage infection but it is clear from the data that a large number of genes are involved in cascades effecting the success of phage infection. Clearly, phage infection is a complicated process involving and relying on a large number of effector genes that can hinder the process effectively when knocked out.

Genes that had a decrease in the abundance of reads at specific insertions compared to the control are those that were associated with phage resistance so it is interesting to note that the gene that had the strongest effect on bacteriophage resistance with a nearly 10 fold decrease in insertions compared to the control was the stringent starvation protein A. Without this gene the bacteriophage was able to infect the cell more easily. This result is in contrast to its reported effects on another bacteriophage, P1 [174]. This is exciting as it indicates that a host gene that might be used to phage advantage by some bacteriophage may also play a resistance role in others. Many of the genes identified are involved in LPS and, specifically, O-antigen synthesis. This is intuitive as the removal of O-antigens would enable bacteriophage easier access to the outer membrane proteins and receptors.

The analysis has focused on fold change in reads for mutants present both with and without phage selection. Those mutants that have been completely eliminated by phage (as evidenced by the sharp drop in unique infections seen in table 5.2) will have 0 insertions in the resistance conferring genes but this will be compared to the number in that gene in the control. This means that the analysis will still be able to identify completely eliminated mutants in the selection verses the control.

The associated genes recognized to be involved in susceptibility and resistance were found in the 4 replicates (10ul phage, 200ul phage, 3 hour and 5 hour time points). A high confidence can be assigned to them as they have effectively been independently replicated 4 times.

It is clear from these selections that susceptibility to bacteriophage infection as well as resistance was dependent on a broad range of genes with different functional characteristics. It would be useful to further characterize and investigate the genes that do not reveal a clear mechanism for disruption of the normal infection pathway. This would be performed in the wet lab by generating gene knockouts and testing the phenotypic effects *in vitro*.

This part of the study highlighted how many factors are able to influence PT and that it is not single gene changes that determine whether a strain is resistant or susceptible to a given bacteriophage. It would be useful to be able to compare the gene lists associated with infection and resistance for different selections using different phages from within the groups of typing phages (chapter 2).

Hypothetically, they would include many of the same genes. However, variations in a small number of specific genes would highlight the differences that determine variation in their infection profiles. It would also be interesting to compare selection results on the same library strain for typing phages from completely different groups to see how many genes vary in the associated infection gene list.

TraDIS has indicated that bacteriophage susceptibility or resistance is not an absolute state but is a balance between the number of genes a strain can have associated with resistance or infection and can be viewed more in the sense of accumulation more to one side of the scale of resistance/susceptibility than the other.

Now that the TraDIS method has been successfully used to look for evidence of bacteriophage survival in susceptible strains, the next steps would be to use the same method but with one of the typing phages that the library strain is resistant to. Hypothetically, more genes with a decrease in insertions < -2 LogFC and less genes associated with infection would be expected.

The genes highlighted by the selection for both infection and resistance can be used as gene markers for resistance or susceptibility for typing phage 13 and, most likely, the other group 2 typing phages but it is clear that they cannot be investigated individually as a single gene that enables infection/resistance. The presence of Stringent starvation protein A in the genome of strain 9000 does not make it resistant to typing phage 13 but the TraDIS results have clearly shown that this gene is contributing to some level of resistance to that phage. Similarly, the presence of OmpC alone is not enough to enable the phage to successfully lyse the cell –translational and transcriptional control is also required. The key differences in phage/host reaction profiles will be revealed when comparing gene lists associated with different typing phage selections. Data from this study has lead to the hypothesis that each typing phage group will produce a slightly different gene list of associated genes and these are the key differences in the host-determined infectivity of the typing phage groups.

5.5 Conclusions

The TraDIS method successfully identified novel regions associated with bacteriophage susceptibility and resistance. The stringent starvation protein A has been implicated in resistance for the first time and the Sap operon has been implicated in infection for the first time. The previous chapters had highlighted the issues with using whole genome sequencing data alone to identify genetic determinants of PT. This is because the accessory genomes of STEC O157 are so large such that even two closely related strains of different PT will have several genes that vary. To be more confident of gene involvement in PT conversion, mutagenesis should be used to provide phenotypic evidence of the gene’s ability to convert a strain. TraDIS is a highly effective whole genome mutagenesis method that facilitates high throughput. During this study TraDIS was shown to be an effective method for investigating the genetic basis of PT.

Chapter 6: Discussion

The genetic determinants of PT for STEC O157 have been a mystery for the past 30 years'; during this PhD I have attempted to unravel the complicated interactions of the typing phages with STEC O157. The aims of this project were to sequence and annotate the 16 phages used to type STEC O157, sequence and annotate strains of STEC O157, analyse the prophage-like elements of STEC O157, identify genetic elements associated with the PT phenotype and build on the current understanding of PT and its representation of clinical significance. The detailed results of this study are described in the previous four chapters and the main findings are summarised below.

The 16 typing phages can be grouped into 4 genetic similarity groups which correlate with their infectivity profiles. This highlighted that the typing scheme could be simplified utilising representatives from each group to give a simpler but less specific PT scheme. It also highlighted that those PTs that vary in their susceptibility within the typing phage groups are likely to be caused by small numbers of genes that vary within the groups. This work formed the bedrock of subsequent analysis because following this analysis it became apparent that common changes in PT (such as PT8 to PT54 or PT32 to PT21/28) are often caused by the strain acquiring resistance to one group of the typing phages.

The emergence of one of the most clinically significant strains in the UK today was caused by the acquisition of Stx2a to the PT32 phenotype to convert it to become PT21/28. This is significant as it provides further evidence of the evolution of the highly pathogenic PT21/28 clone from the PT32 strain following the acquisition of the Stx2a phage. It is now clear that the rapid emergence of this clone 25 years ago was caused by the acquisition of the Stx2a phage. I was also able to show that the Stx2a phage metabolically repressed its host and therefore may have further adapted it to the bovine and/or human host and that this may also have contributed towards the success of this clone in clinical cases. I suggest that the PT21/28 clone is clinically significant in the UK because of the acquisition of the Stx2a prophage that conveys bacteriophage resistance, pathogenicity and metabolic adaptation.

Long read sequencing revealed the high rate of microevolution of STEC O157 within a short period of time and the ability of an IncHI2 plasmid to convert a PT8 strain to PT54. This built on the current understanding of PT in terms of its representation of clinical significance by implicating the acquisition of the IncHI2 plasmid to the PT54 phenotype. The IncHI2 plasmid has multiple clinical consequences because it conveys antibiotic resistance and also has been shown to confer increased acid and temperature resistance. This means that the PT54 phenotype that is conveyed by the IncHI2 plasmid also goes hand in hand with antibiotic resistance and gut survival adaptation in differing conditions. This may mean that there is an increased likelihood of human-to-human transmission.

TraDIS sequencing implicated 114 genes in phage infectivity and 44 genes in phage resistance in STEC O157. Interestingly, even though the 44 genes identified in phage resistance were mostly cell wall or O antigen associated, the 114 genes identified in phage infectivity were from a broad range of COG assignments indicating that phage infection requires the involvement of multiple pathways for successful lysis. This is further evidence that the determination of PT is decided by multiple factors and is likely to not always be determined by the presence or absence of just one gene. The TraDIS project proved the most useful in terms of being high-throughput for finding genetic candidates for the determinants of PT and ultimately was probably the best method to answer the original questions of the thesis.

It is intuitive that the reason why we see some PTs emerging and proliferating more widely than others is because the PT is a marker of clinical significance or pathoadaptation. Therefore, the genes that determine a particular PT may be linked to genes that contribute to the ability of the strain to survive in the environment or to become pathoadapted to a specific host. Genes associated with changes in PT and those associated with enhanced survivial or pathoadaptation may be carried on mobile elements like prophages or plasmids and this is precisely what I have found for two PTs – PT21/28 and PT54. The two PTs that account for 61% of the reported STEC O157 cases found in the UK are PT21/28 which is associated with domestic infection and PT8 from which a subset is associated with foreign travel and a subset is domestic. I propose that the reason for the success of the PT21/28 clone and its PT phenotype is due to a shiga toxin phage mobile element. The success of the other dominant PT (PT8) has the potential to increase even further with the addition of the IncHI2 plasmid mobile element which conveys survival advantages to the host. As the use of antibiotics continues we may see further spread of the IncHI2 plasmid in more foreign travel associated STEC O157 cases and the IncHI2 plasmid may convert more representatives of the PT8 clone to PT54 to become the dominant foreign travel associated PT. From this work we can conclude that those mobile elements associated with dominant PT conversion are likely to also carry evolutionary or pathoadaptivity advantages.

The fact that each PT is produced by a profile of lysis of 16 TPs means that the resulting PT phenotype is from the layering of multiple mechanisms producing resistance and susceptibility to each of the four TP groups. It is much easier to study the difference between two closely related PTs like PT8 and PT54 (that only differ in their reaction to the group 3 TPs) than to compare two unrelated PTs to try and find the underlying genetic mechanisms. It is easier to consider each PT as the result of four different mechanisms accounting for the resistance or susceptibility seen for each TP group. However some mechanisms may apply to more than one group of the TPs.

Long read sequencing would enable a pangenome analysis to draw out associated genes for resistance or susceptibility for each TP group, unfortunately this is not possible with short read data because it is likely that many of the associated genes will be mobile element associated and therefore not resolved very well by short reads. For the statistical significance needed for genome-wide association studies long read sequences of a very high number of strains would be required. Currently, this is still very expensive, but would ultimately provide a whole genome scanning effect.

Whole genome scanning was used in the TraDIS analysis and was successful in identifying several genes that alter TP13 susceptibility or resistance significantly. Some of these genes are likely to only be applicable to resistance and susceptibility of TP group 2 but there are also likely to be more generic phage resistance mechanisms like LPS associated genes. It is important to highlight that TraDIS implicates all the possible individual genes that can be used to render a strain resistant or susceptible. It is likely that if you compare two closely related strains that have a different susceptibility to TP13 that they will differ in just one of these genes.

Bacteriophage and bacterial co-evolution in the environment is likely to be a significant driver for the range of different PTs that we see in the UK. The most common PTs in the UK are 2, 8, 21/28 and 32 and they were all found in different modules during the modularity analysis. This is significant because it is evidence of the different niches that phages create in the environment so that resistances to subsets of typing phages is evolutionarily advantageous. However, the competition of phage evolution is so great that resistance to all is relatively rare even though completely resistant strains are seen in the UK. Each dominant PT has its own module that is indicative of the phages that it encounters in its environmental niche but would be less likely to survive in environments exposing it to phages seen in the other modules or environmental niches. This implies that we are sampling 4 different potential niches or susceptibility profiles that will better enable the survival of STEC O157 in the environment with regards to the current generations of bacteriophage.

The role of phage typing as a reference microbiology typing method is outdated as more reliable, robust and evolutionarily informative sequence based methods emerge. This work has highlighted the role that mobile elements play in phage typing which can confound outbreak detection and investigation. However the fact that PT is linked to mobile elements has also revealed why certain phage types have associated pathogenicity or environmental niches.

This work has given multiple insights into the complex relationships of bacteria and bacteriophage and will provide insights into best methods for analysis for future projects. I have established that short read sequencing is insufficient and long read sequencing is a better way to use sequencing data to determine PT. However, it has become clear that the most effective way to investigate phage susceptibility or resistance is through TraDIS.

6.1 Further work

This work could be taken further by greater characterisation of single genes involved in PT. This would involve characterising which gene/s on the IncHI2 plasmid were responsible for the change from PT8 to PT54, which gene/s on the Stx2a prophage region were responsible for the change from PT32 to PT21/28 and confirming TraDIS implicated gene/s converting a strain's susceptibility or resistance to TP13. This can be done with specific genetic knockouts in the lab and would confirm these genes as genetic markers for that particular PT associated phenotype.

In terms of recognising new genetic determinants of PT:

- Machine learning techniques could be used to train the computer to trawl large amounts of sequencing data to find statistically implicated genes associated with PT. Machine learning has been used for many studies that use artificial neural networks and random forests [175, 176]. Classifiers can be created based on genomic features of known sets of phage-bacterium interactions, as defined in this thesis, and could be trialed on test sets for which interaction results are 'unknown'. Once the model has been trained on the data it can be used on new data to find new classifiers and new genetic determinants of PT.
- Pangenome analysis, as already mentioned, would similarly be a high throughput method for genome wide association and has been used successfully by many people to implicate genes in specific phenotypes so long as the metadata is provided, a high number of strains are used and strong statistical analysis is employed [177, 178].
- More TraDIS experiments should be performed using the other typing phages as the selection to produce implicated gene lists for the other

typing phage groups. This method should also be applied to more bacteria-phage interactions and especially those that are MDR. To improve our arsenal against MDR bacteria we should employ phage therapy but further work needs to be done to understand the bacteria-phage interactions for some of the MDR bacteria. TraDIS is a highly valuable method to investigate this.

6.2 Limitations

One of the shortcomings of the work is that only a limited number of interactions that determine PT have been discovered. This means that genetic determinants for all the PTs cannot be implemented into an *In silico* phage typing scheme. Another shortcoming is that only one typing phage was able to be tested on the TraDIS library. This is due to time constraints, the TraDIS method was only decided on towards the end of year 1 and library production took ~8 months. Better planning at the beginning of this PhD may have meant that I was able spend more time on the TraDIS library and less time trying to analyse short read sequencing which ultimately yielded very little. The Stx2a and PT21/28 link was an important finding that should have been investigated far enough to determine the gene present on the prophage that caused the PT switch. This could have also easily been achieved with more dedicated time. Some may argue that more of a lab based approach should have been used to investigate PT but this was not in the original aims of the study so, although lab-based methods were also employed, whole genome sequencing methods were used in all aspects of the investigation. Others would also argue that the *E.coli* bacteriophage receptors are relatively well studied and the TraDIS results regarding OmpC and it's regulators were not novel. I agree with this but think that the value of TraDIS is that we were able to recognise 113 other genes that play a role in bacteriophage infection.

In reality, the public health significance of PT is better investigated by the associated virulence factors than by the specific genes that convey resistance or susceptibility to the typing phages. These are likely to be co-acquired or could even be prior to the genes that determine current PT. However this body of work has shown that they can also be co-acquired in single genetic events such as plasmid or prophage gain.

6.3 Conclusions

STEC O157 poses a significant public health threat in the UK with severe clinical complications associated with certain strains. This study showed that the virulence of certain strains can be attributed to mobile elements that also convey bacteriophage resistance therefore providing a survival advantage. This work has also shown that the epidemiologically highlighted dominant clones of STEC O157 are occupying different niches in terms of bacteriophage survival in the environment. It has highlighted the complexity of the genetic determinants of PT but proven that there are methods to untangle this using genome wide mutagenesis.

Reference List

1. Cukrowska B, Lodinova-Zadnikova R, Enders C, Sonnenborn U, Schulze J, Tlaskalova-Hogenova H: Specific proliferative and antibody responses of premature infants to intestinal colonization with nonpathogenic probiotic E. coli strain Nissle 1917. *Scand J Immunol* 2002, 55:204-209.

2. Kaper JB, Nataro JP, Mobley HL: Pathogenic Escherichia coli. *Nat Rev Microbiol* 2004, 2:123-140.

3. Ferens WA, Hovde CJ: Escherichia coli O157:H7: animal reservoir and sources of human infection. *Foodborne Pathog Dis* 2011, 8:465-487.

4. Tuttle J, Gomez T, Doyle MP, Wells JG, Zhao T, Tauxe RV, Griffin PM: Lessons from a large outbreak of Escherichia coli O157:H7 infections: insights into the infectious dose and method of widespread contamination of hamburger patties. *Epidemiol Infect* 1999, 122:185-192.

5. Feng P, Lampel KA, Karch H, Whittam TS: Genotypic and phenotypic changes in the emergence of Escherichia coli O157:H7. *J Infect Dis* 1998, 177:1750-1753.

6. Dallman TJ, Ashton PM, Byrne L, Perry NT, Petrovska L, Ellis R, Allison L, Hanson MF, Holmes A, Gunn GJ, Chase-Topping ME, Woolhouse ME, Grant KA, Gally DL, Wain J, Jenkins C: Applying phylogenomics to understand the emergence of Shiga-toxin-producing Escherichia coli O157:H7 strains causing severe human disease in the UK. *Microbial Genomics* 2015.

7. Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C, Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, Shinagawa H: Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res* 2001, 8:11-22.

8. Ashton PM, Perry N, Ellis R, Petrovska L, Wain J, Grant KA, Jenkins C, Dallman TJ: Insight into Shiga toxin genes encoded by Escherichia coli O157 from whole genome sequencing. *PeerJ* 2015, 3:e739.

9. Scheutz F, Teel LD, Beutin L, Pierard D, Buvens G, Karch H, Mellmann A, Caprioli A, Tozzoli R, Morabito S, Strockbine NA, Melton-Celsa AR, Sanchez M, Persson S, O'Brien AD: Multicenter evaluation of a sequence-based protocol for subtyping Shiga toxins and standardizing Stx nomenclature. *J Clin Microbiol* 2012, 50:2951-2963.

10. Pennington H: Escherichia coli O157. *Lancet* 2010, 376:1428-1435.

11. Hagel C, Krasemann S, Loffler J, Puschel K, Magnus T, Glatzel M: Upregulation of Shiga toxin receptor CD77/Gb3 and interleukin-1beta expression in the

brain of EHEC patients with hemolytic uremic syndrome and neurologic symptoms. *Brain Pathol* 2015, 25:146-156.

12. Pruimboom-Brees IM, Morgan TW, Ackermann MR, Nystrom ED, Samuel JE, Cornick NA, Moon HW: Cattle lack vascular receptors for Escherichia coli O157:H7 Shiga toxins. *Proc Natl Acad Sci U S A* 2000, 97:10325-10329.

13. James CE, Stanley KN, Allison HE, Flint HJ, Stewart CS, Sharp RJ, Saunders JR, McCarthy AJ: Lytic and lysogenic infection of diverse Escherichia coli and Shigella strains with a verocytotoxigenic bacteriophage. *Appl Environ Microbiol* 2001, 67:4335-4337.

14. McDaniel TK, Jarvis KG, Donnenberg MS, Kaper JB: A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens. *Proc Natl Acad Sci U S A* 1995, 92:1664-1668.

15. Franzin FM, Sircili MP: Locus of enterocyte effacement: a pathogenicity island involved in the virulence of enteropathogenic and enterohemorragic Escherichia coli subjected to a complex network of gene regulation. *Biomed Res Int* 2015, 2015:534738.

16. Knutton S, Rosenshine I, Pallen MJ, Nisan I, Neves BC, Bain C, Wolff C, Dougan G, Frankel G: A novel EspA-associated surface organelle of enteropathogenic Escherichia coli involved in protein translocation into epithelial cells. *EMBO J* 1998, 17:2166-2176.

17. Ide T, Laarmann S, Greune L, Schillers H, Oberleithner H, Schmidt MA: Characterization of translocation pores inserted into plasma membranes by type III-secreted Esp proteins of enteropathogenic Escherichia coli. *Cell Microbiol* 2001, 3:669-679.

18. Kenny B, Devinney R, Stein M, Reinscheid DJ, Frey EA, Finlay BB: Enteropathogenic E. coli (EPEC) transfers its receptor for intimate adherence into mammalian cells. *Cell* 1997, 91:511-520.

19. Xu X, McAteer SP, Tree JJ, Shaw DJ, Wolfson EB, Beatson SA, Roe AJ, Allison LJ, Chase-Topping ME, Mahajan A, Tozzoli R, Woolhouse ME, Morabito S, Gally DL: Lysogeny with Shiga toxin 2-encoding bacteriophages represses type III secretion in enterohemorrhagic Escherichia coli. *PLoS Pathog* 2012, 8:e1002672.

20. Tree JJ, Roe AJ, Flockhart A, McAteer SP, Xu X, Shaw D, Mahajan A, Beatson SA, Best A, Lotz S, Woodward MJ, La RR, Murphy KC, Leong JM, Gally DL: Transcriptional regulators of the GAD acid stress island are carried by effector protein-encoding prophages and indirectly control type III secretion in enterohemorrhagic Escherichia coli O157:H7. *Mol Microbiol* 2011, 80:1349-1365.

21. Dame RT, Noom MC, Wuite GJ: Bacterial chromatin organization by H-NS protein unravelled using dual DNA manipulation. *Nature* 2006, 444:387-390.

22. Gould LH, Demma L, Jones TF, Hurd S, Vugia DJ, Smith K, Shiferaw B, Segler S, Palmer A, Zansky S, Griffin PM: Hemolytic uremic syndrome and death in persons with Escherichia coli O157:H7 infection, foodborne diseases active surveillance network sites, 2000-2006. *Clin Infect Dis* 2009, 49:1480-1485.

23. Byrne L, Jenkins C, Launders N, Elson R, Adak GK: The epidemiology, microbiology and clinical impact of Shiga toxin-producing Escherichia coli in England, 2009-2012. *Epidemiol Infect* 2015, 143:3475-3487.

24. Underwood AP, Dallman T, Thomson NR, Williams M, Harker K, Perry N, Adak B, Willshaw G, Cheasty T, Green J, Dougan G, Parkhill J, Wain J: Public health value of next-generation DNA sequencing of enterohemorrhagic Escherichia coli isolates from an outbreak. *J Clin Microbiol* 2013, 51:232-237.

25. Chase-Topping M, Gally D, Low C, Matthews L, Woolhouse M: Super-shedding and the link between human infection and livestock carriage of Escherichia coli O157. *Nat Rev Microbiol* 2008, 6:904-912.

26. Low JC, McKendrick IJ, Mckechnie C, Fenlon D, Naylor SW, Currie C, Smith DG, Allison L, Gally DL: Rectal carriage of enterohemorrhagic Escherichia coli O157 in slaughtered cattle. *Appl Environ Microbiol* 2005, 71:93-97.

27. Naylor SW, Low JC, Besser TE, Mahajan A, Gunn GJ, Pearce MC, McKendrick IJ, Smith DG, Gally DL: Lymphoid follicle-dense mucosa at the terminal rectum is the principal site of colonization of enterohemorrhagic Escherichia coli O157:H7 in the bovine host. *Infect Immun* 2003, 71:1505-1512.

28. Tarr PI, Gordon CA, Chandler WL: Shiga-toxin-producing Escherichia coli and haemolytic uraemic syndrome. *Lancet* 2005, 365:1073-1086.

29. Bayliss L, Carr R, Edeghere O, Knapper E, Nye K, Harvey G, Adak G, Duggal H: School outbreak of Escherichia coli O157 with high levels of transmission, Staffordshire, England, February 2012. *J Public Health (Oxf)* 2015.

30. Gillespie IA, O'Brien SJ, Adak GK, Cheasty T, Willshaw G: Foodborne general outbreaks of Shiga toxin-producing Escherichia coli O157 in England and Wales 1992-2002: where are the risks? *Epidemiol Infect* 2005, 133:803-808.

31. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M: Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 1977, 265:687-695.

32. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP,

Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM: Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005, 437:376-380.

33. Jenkins C, Dallman TJ, Launders N, Willis C, Byrne L, Jorgensen F, Eppinger M, Adak GK, Aird H, Elviss N, Grant KA, Morgan D, McLauchlin J: Public Health Investigation of Two Outbreaks of Shiga Toxin-Producing Escherichia coli O157 Associated with Consumption of Watercress. *Appl Environ Microbiol* 2015, 81:3946-3952.

34. Dallman TJ, Byrne L, Launders N, Glen K, Grant KA, Jenkins C: The utility and public health implications of PCR and whole genome sequencing for the detection and investigation of an outbreak of Shiga toxin-producing Escherichia coli serogroup O26:H11. *Epidemiol Infect* 2015, 143:1672-1680.

35. Inns T, Lane C, Peters T, Dallman T, Chatt C, McFarland N, Crook P, Bishop T, Edge J, Hawker J, Elson R, Neal K, Adak GK, Cleary P: A multi-country Salmonella Enteritidis phage type 14b outbreak associated with eggs from a German producer: 'near real-time' application of whole genome sequencing and food chain investigations, United Kingdom, May to September 2014. *Euro Surveill* 2015, 20.

36. Koser CU, Ellington MJ, Cartwright EJ, Gillespie SH, Brown NM, Farrington M, Holden MT, Dougan G, Bentley SD, Parkhill J, Peacock SJ: Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathog* 2012, 8:e1002824.

37. Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT, Adak G, Petrovska L, Ellis RJ, Elson R, Underwood A, Green J, Hanage WP, Jenkins C, Grant K, Wain J: Whole-Genome Sequencing for National Surveillance of Shiga Toxin-Producing Escherichia coli O157. *Clin Infect Dis* 2015, 61:305-312.

38. Bashir A, Klammer AA, Robins WP, Chin CS, Webster D, Paxinos E, Hsu D, Ashby M, Wang S, Peluso P, Sebra R, Sorenson J, Bullard J, Yen J, Valdovino M, Mollova E, Luong K, Lin S, LaMay B, Joshi A, Rowe L, Frace M, Tarr CL, Turnsek M, Davis BM, Kasarskis A, Mekalanos JJ, Waldor MK, Schadt EE: A hybrid approach for the automated finishing of bacterial genomes. *Nat Biotechnol* 2012, 30:701-707.

39. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A, Ouedraogo N, Afrough B, Bah A, Baum JH, Becker-Ziaja B, Boettcher JP, Cabeza-Cabrerizo M, Camino-Sanchez A, Carter LL, Doerrbecker J, Enkirch T, Garcia-Dorival I, Hetzelt N, Hinzmann J, Holm T, Kafetzopoulou LE, Koropogui M, Kosgey A, Kuisma E, Logue CH, Mazzarelli A, Meisel S, Mertens M, Michel J, Ngabo D, Nitzsche K, Pallasch E, Patrono LV, Portmann J, Repits JG, Rickett NY, Sachse A, Singethan K, Vitoriano I, Yemanaberhan RL, Zekeng EG, Racine T, Bello A, Sall AA, Faye

O, Faye O, Magassouba N, Williams CV, Amburgey V, Winona L, Davis E, Gerlach J, Washington F, Monteil V, Jourdain M, Bererd M, Camara A, Somlare H, Camara A, Gerard M, Bado G, Baillet B, Delaune D, Nebie KY, Diarra A, Savane Y, Pallawo RB, Gutierrez GJ, Milhano N, Roger I, Williams CJ, Yattara F, Lewandowski K, Taylor J, Rachwal P, Turner DJ, Pollakis G, Hiscox JA, Matthews DA, O'Shea MK, Johnston AM, Wilson D, Hutley E, Smit E, Di CA, Wolfel R, Stoecker K, Fleischmann E, Gabriel M, Weller SA, Koivogui L, Diallo B, Keita S, Rambaut A, Formenty P, Gunther S, Carroll MW: Real-time, portable genome sequencing for Ebola surveillance. *Nature* 2016, 530:228-232.

40. Byrne L, Elson R, Dallman TJ, Perry N, Ashton P, Wain J, Adak GK, Grant KA, Jenkins C: 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* 2014, 9:e85901.

41. Ahmed R, Bopp C, Borczyk A, Kasatiya S: Phage-typing scheme for Escherichia coli O157:H7. *J Infect Dis* 1987, 155:806-809.

42. Schurch AC, van SD: DNA fingerprinting of Mycobacterium tuberculosis: from phage typing to whole-genome sequencing. *Infect Genet Evol* 2012, 12:602-609.

43. Baggesen DL, Sorensen G, Nielsen EM, Wegener HC: Phage typing of Salmonella Typhimurium - is it still a useful tool for surveillance and outbreak investigation? *Euro Surveill* 2010, 15:19471.

44. Turbadkar SD, Ghadge DP, Patil S, Chowdhary AS, Bharadwaj R: Circulating phage type of Vibrio cholerae in Mumbai. *Indian J Med Microbiol* 2007, 25:177-178.

45. Hausmann R: Bacteriophage T7 genetics. *Curr Top Microbiol Immunol* 1976, 75:77-110.

46. Miller ES, Kutter E, Mosig G, Arisaka F, Kunisawa T, Ruger W: Bacteriophage T4 genome. *Microbiol Mol Biol Rev* 2003, 67:86-156, table.

47. Serwer P, Wright ET, Hakala KW, Weintraub ST: Evidence for bacteriophage T7 tail extension during DNA injection. *BMC Res Notes* 2008, 1:36.

48. Weitz JS, Poisot T, Meyer JR, Flores CO, Valverde S, Sullivan MB, Hochberg ME: Phage-bacteria infection networks. *Trends Microbiol* 2013, 21:82-91.

49. Adams NL, Byrne L, Smith GA, Elson R, Harris JP, Salmon R, Smith R, O'Brien SJ, Adak GK, Jenkins C: Shiga Toxin-Producing Escherichia coli O157, England and Wales, 1983-2012

35. *Emerg Infect Dis* 2016, 22:590-597.

50. Rakhuba DV, Kolomiets EI, Dey ES, Novik GI: Bacteriophage receptors, mechanisms of phage adsorption and penetration into host cell. *Pol J Microbiol* 2010, 59:145-155.

51. Mark DF, Richardson CC: Escherichia coli thioredoxin: a subunit of bacteriophage T7 DNA polymerase. *Proc Natl Acad Sci U S A* 1976, 73:780-784.

52. Broussard GW, Hatfull GF: Evolution of genetic switch complexity. *Bacteriophage* 2013, 3:e24186.

53. Campbell AM: Chromosomal insertion sites for phages and plasmids. *J Bacteriol* 1992, 174:7495-7499.

54. Mizuuchi K, Kemper B, Hays J, Weisberg RA: T4 endonuclease VII cleaves holliday structures. *Cell* 1982, 29:357-365.

55. Kreuzer KN, Brister JR: Initiation of bacteriophage T4 DNA replication and replication fork dynamics: a review in the Virology Journal series on bacteriophage T4 and its relatives. *Virol J* 2010, 7:358.

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

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

58. Lederberg EM, Lederberg J: Genetic Studies of Lysogenicity in Escherichia Coli. *Genetics* 1953, 38:51-64.

59. Cumby N, Davidson AR, Maxwell KL: The moron comes of age. *Bacteriophage* 2012, 2:225-228.

60. Hendrix RW: Bacteriophage genomics. *Curr Opin Microbiol* 2003, 6:506-511.

61. Ogura Y, Mondal SI, Islam MR, Mako T, Arisawa K, Katsura K, Ooka T, Gotoh Y, Murase K, Ohnishi M, Hayashi T: The Shiga toxin 2 production level in enterohemorrhagic Escherichia coli O157:H7 is correlated with the subtypes of toxin-encoding phage. *Sci Rep* 2015, 5:16663.

62. Young R: Bacteriophage lysis: mechanism and regulation. *Microbiol Rev* 1992, 56:430-481.

63. Streisinger G, Mukai F, DREYER WJ, Miller B, HORIUCHI S: Mutations affecting the lysozyme of phage T4. *Cold Spring Harb Symp Quant Biol* 1961, 26:25-30.

64. Kao SH, McClain WH: Baseplate protein of bacteriophage T4 with both structural and lytic functions. *J Virol* 1980, 34:95-103.

65. Visconti N: Resistance to lysis from without in bacteria infected with T2 bacteriophage. *J Bacteriol* 1953, 66:247-253.

66. Bode W: Lysis inhibition in Escherichia coli infected with bacteriophage T4. *J Virol* 1967, 1:948-955.

67. Abedon ST: Bacteriophage secondary infection. *Virol Sin* 2015, 30:3-10.

68. Josslin R: The lysis mechanism of phage T4: mutants affecting lysis. *Virology* 1970, 40:719-726.

69. Rennell D, Poteete AR: Phage P22 lysis genes: nucleotide sequences and functional relationships with T4 and lambda genes. *Virology* 1985, 143:280-289.

70. Champe SP: BACTERIOPHAGE REPRODUCTION. *Annu Rev Microbiol* 1963, 17:87-114.

71. Hoskisson PA, Smith MC: Hypervariation and phase variation in the bacteriophage 'resistome'. *Curr Opin Microbiol* 2007, 10:396-400.

72. Seed KD, Lazinski DW, Calderwood SB, Camilli A: A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* 2013, 494:489-491.

73. Labrie SJ, Samson JE, Moineau S: Bacteriophage resistance mechanisms. *Nat Rev Microbiol* 2010, 8:317-327.

74. Perry LL, SanMiguel P, Minocha U, Terekhov AI, Shroyer ML, Farris LA, Bright N, Reuhs BL, Applegate BM: Sequence analysis of Escherichia coli O157:H7 bacteriophage PhiV10 and identification of a phage-encoded immunity protein that modifies the O157 antigen. *FEMS Microbiol Lett* 2009, 292:182-186.

75. Destoumieux-Garzon D, Duquesne S, Peduzzi J, Goulard C, Desmadril M, Letellier L, Rebuffat S, Boulanger P: The iron-siderophore transporter FhuA is the receptor for the antimicrobial peptide microcin J25: role of the microcin Val11-Pro16 beta-hairpin region in the recognition mechanism. *Biochem J* 2005, 389:869-876.

76. Comeau AM, Krisch HM: War is peace--dispatches from the bacterial and phage killing fields. *Curr Opin Microbiol* 2005, 8:488-494.

77. Tock MR, Dryden DT: The biology of restriction and anti-restriction. *Curr Opin Microbiol* 2005, 8:466-472.

78. Meisel A, Bickle TA, Kruger DH, Schroeder C: Type III restriction enzymes need two inversely oriented recognition sites for DNA cleavage. *Nature* 1992, 355:467-469.

79. Kruger DH, Schroeder C, Santibanez-Koref M, Reuter M: Avoidance of DNA methylation. A virus-encoded methylase inhibitor and evidence for counterselection of methylase recognition sites in viral genomes. *Cell Biophys* 1989, 15:87-95.

80. Sulakvelidze A, Alavidze Z, Morris JG, Jr.: Bacteriophage therapy. *Antimicrob Agents Chemother* 2001, 45:649-659.

81. Young R, Gill JJ: MICROBIOLOGY. Phage therapy redux--What is to be done? *Science* 2015, 350:1163-1164.

82. Miedzybrodzki R, Fortuna W, Weber-Dabrowska B, Gorski A: Phage therapy of staphylococcal infections (including MRSA) may be less expensive than antibiotic treatment. *Postepy Hig Med Dosw (Online)* 2007, 61:461-465.

83. Sarker SA, Sultana S, Reuteler G, Moine D, Descombes P, Charton F, Bourdin G, McCallin S, Ngom-Bru C, Neville T, Akter M, Huq S, Qadri F, Talukdar K, Kassam M, Delley M, Loiseau C, Deng Y, El AS, Berger B, Brussow H: Oral Phage Therapy of Acute Bacterial Diarrhea With Two Coliphage Preparations: A Randomized Trial in Children From Bangladesh. *EBioMedicine* 2016, 4:124-137.

84. Leung SS, Parumasivam T, Gao FG, Carrigy NB, Vehring R, Finlay WH, Morales S, Britton WJ, Kutter E, Chan HK: Production of Inhalation Phage Powders Using Spray Freeze Drying and Spray Drying Techniques for Treatment of Respiratory Infections. *Pharm Res* 2016.

85. Wagenaar JA, van Bergen MA, Mueller MA, Wassenaar TM, Carlton RM: Phage therapy reduces Campylobacter jejuni colonization in broilers. *Vet Microbiol* 2005, 109:275-283.

86. Chao MC, Abel S, Davis BM, Waldor MK: The design and analysis of transposon insertion sequencing experiments. *Nat Rev Microbiol* 2016, 14:119-128.

87. Langridge GC, Phan MD, Turner DJ, Perkins TT, Parts L, Haase J, Charles I, Maskell DJ, Peters SE, Dougan G, Wain J, Parkhill J, Turner AK: Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome Res* 2009, 19:2308-2316.

88. Zhang YJ, Ioerger TR, Huttenhower C, Long JE, Sassetti CM, Sacchettini JC, Rubin EJ: Global assessment of genomic regions required for growth in Mycobacterium tuberculosis. *PLoS Pathog* 2012, 8:e1002946.

89. Pickard D, Kingsley RA, Hale C, Turner K, Sivaraman K, Wetter M, Langridge G, Dougan G: A genomewide mutagenesis screen identifies multiple genes contributing to Vi capsular expression in Salmonella enterica serovar Typhi. *J Bacteriol* 2013, 195:1320-1326.

90. Kropinski AM, Lingohr EJ, Moyles DM, Chibeu A, Mazzocco A, Franklin K, Villegas A, Ahmed R, She YM, Johnson RP: Escherichia coli O157:H7 typing phage V7 is a T4-like virus. *J Virol* 2012, 86:10246.

91. Kropinski AM, Waddell T, Meng J, Franklin K, Ackermann HW, Ahmed R, Mazzocco A, Yates J, III, Lingohr EJ, Johnson RP: The host-range, genomics and proteomics of Escherichia coli O157:H7 bacteriophage rV5. *Virol J* 2013, 10:76.

92. Santos MA: An improved method for the small scale preparation of bacteriophage DNA based on phage precipitation by zinc chloride. *Nucleic Acids Res* 1991, 19:5442.

93. Zerbino DR, Birney E: Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 2008, 18:821-829.

94. Seemann T: Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014.

95. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA: SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012, 19:455-477.

96. McCune B, Grace JB: Analysis of Ecological Communities. *MjM Software, Gleneden Beach, Ore* 2002.

97. Liu X, Murata T: An efficient Algortihm for Optimizing Bipartite Modularity in Bipartite Networks. *JACIII* 2010, 14:408-415.

98. Beckett SJ: A weighted modularity algorithm for bipartite networks. *figshare http://dx doi org/10 6084/m9 figshare 999114* 2014.

99. Beckett SJ, Boulton CA, Williams HT: FALCON: a software package for analysis of nestedness in bipartite networks. *F1000Res* 2014, 3:185.

100. Almeida-Neto M, Guimaraes P, Guimaraes Jr PR, Loyola RD, Ulrich W: A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos* 2008, 117:1227-1239.

101. Atmar W, Patterson BD: The measure of order and disorder in the distribution of species in fragmented habitat. *Oecologia* 1993, 96:373-382.

102. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H: vegan: Community Ecology Package. *R package version 2 0-10 http://CRAN R-project org/package=vegan* 2013.

103. Brualdi RA, Sanderson JG: Nested species subsets, gaps, and discrepancy. *Oecologia* 1999, 119:256-264.

104. Bascompte J, Patterson BD: The nested assembly of plant-animal mutualistic networks. *Proc Natl Acad Sci U S A* 1993, 100:9383-0387.

105. Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA: BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 2011, 12:402.

106. Sullivan MJ, Petty NK, Beatson SA: Easyfig: a genome comparison visualizer. *Bioinformatics* 2011, 27:1009-1010.

107. **Riede I, Degen M, Henning U: The receptor specificity of bacteriophages can be determined by a tail fiber modifying protein. *EMBO J* 1985, 4:2343-2346.**

108. **Matthews L, Reeve R, Gally DL, Low JC, Woolhouse ME, McAteer SP, Locking ME, Chase-Topping ME, Haydon DT, Allison LJ, Hanson MF, Gunn GJ, Reid SW: Predicting the public health benefit of vaccinating cattle against Escherichia coli O157. *Proc Natl Acad Sci U S A* 2013, 110:16265-16270.**

109. **Leiman PG, Battisti AJ, Bowman VD, Stummeyer K, Muhlenhoff M, Gerardy-Schahn R, Scholl D, Molineux IJ: The structures of bacteriophages K1E and K1-5 explain processive degradation of polysaccharide capsules and evolution of new host specificities. *J Mol Biol* 2007, 371:836-849.**

110. **Hattman S, FUKASAWA T: HOST-INDUCED MODIFICATION OF T-EVEN PHAGES DUE TO DEFECTIVE GLUCOSYLATION OF THEIR DNA. *Proc Natl Acad Sci U S A* 1963, 50:297-300.**

111. **Springman R, Badgett MR, Molineux IJ, Bull JJ: Gene order constrains adaptation in bacteriophage T7. *Virology* 2005, 341:141-152.**

112. **Beckett SJ, Williams HT: Coevolutionary diversification creates nested-modular structure in phage-bacteria interaction networks. *Interface Focus* 2013, 3:20130033.**

113. **Flores CO, Meyer JR, Valverde S, Farr L, Weitz JS: Statistical structure of host-phage interactions. *Proc Natl Acad Sci U S A* 2011, 108:E288-E297.**

114. **Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasic L, Thingstad TF, Rohwer F, Mira A: Explaining microbial population genomics through phage predation. *Nat Rev Microbiol* 2009, 7:828-836.**

115. **Mora A, Blanco M, Blanco JE, Alonso MP, Dhabi G, Thomson-Carter F, Usera MA, Bartolome R, Prats G, Blanco J: Phage types and genotypes of shiga toxin-producing Escherichia coli O157:H7 isolates from humans and animals in spain: identification and characterization of two predominating phage types (PT2 and PT8). *J Clin Microbiol* 2004, 42:4007-4015.**

116. **Watson N: A new revision of the sequence of plasmid pBR322. *Gene* 1988, 70:399-403.**

117. **Merlin C, McAteer S, Masters M: Tools for characterization of Escherichia coli genes of unknown function. *J Bacteriol* 2002, 184:4573-4581.**

118. **Darling AE, Mau B, Perna NT: progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 2010, 5:e11147.**

119. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990, 215:403-410.**

120. **Rose RE: The nucleotide sequence of pACYC184. *Nucleic Acids Res* 1988, 16:355.**

121. Wang D, Zetterstrom CE, Gabrielsen M, Beckham KS, Tree JJ, Macdonald SE, Byron O, Mitchell TJ, Gally DL, Herzyk P, Mahajan A, Uvell H, Burchmore R, Smith BO, Elofsson M, Roe AJ: Identification of bacterial target proteins for the salicylidene acylhydrazide class of virulence-blocking compounds. *J Biol Chem* 2011, 286:29922-29931.

122. Vaas LA, Sikorski J, Hofner B, Fiebig A, Buddruhs N, Klenk HP, Goker M: opm: an R package for analysing OmniLog(R) phenotype microarray data. *Bioinformatics* 2013, 29:1823-1824.

123. Matthews L, Low JC, Gally DL, Pearce MC, Mellor DJ, Heesterbeek JA, Chase-Topping M, Naylor SW, Shaw DJ, Reid SW, Gunn GJ, Woolhouse ME: Heterogeneous shedding of Escherichia coli O157 in cattle and its implications for control. *Proc Natl Acad Sci U S A* 2006, 103:547-552.

124. Dodd IB, Perkins AJ, Tsemitsidis D, Egan JB: Octamerization of lambda CI repressor is needed for effective repression of P(RM) and efficient switching from lysogeny. *Genes Dev* 2001, 15:3013-3022.

125. KAISER AD, JACOB F: Recombination between related temperate bacteriophages and the genetic control of immunity and prophage localization. *Virology* 1957, 4:509-521.

126. Mauro SA, Koudelka GB: Shiga toxin: expression, distribution, and its role in the environment
61. *Toxins (Basel)* 2011, 3:608-625.

127. Ohlendorf DH, Tronrud DE, Matthews BW: Refined structure of Cro repressor protein from bacteriophage lambda suggests both flexibility and plasticity. *J Mol Biol* 1998, 280:129-136.

128. Takeda Y, Folkmanis A, Echols H: Cro regulatory protein specified by bacteriophage lambda. Structure, DNA-binding, and repression of RNA synthesis. *J Biol Chem* 1977, 252:6177-6183.

129. Zinder ND: Lysogenization and superinfection immunity in Salmonella. *Virology* 1958, 5:291-326.

130. Veses-Garcia M, Liu X, Rigden DJ, Kenny JG, McCarthy AJ, Allison HE: Transcriptomic analysis of Shiga-toxigenic bacteriophage carriage reveals a profound regulatory effect on acid resistance in Escherichia coli. *Appl Environ Microbiol* 2015, 81:8118-8125.

131. Cornick NA, Helgerson AF, Mai V, Ritchie JM, Acheson DW: In vivo transduction of an Stx-encoding phage in ruminants
47. *Appl Environ Microbiol* 2006, 72:5086-5088.

132. Muniesa M, Jofre J: Occurrence of phages infecting Escherichia coli O157:H7 carrying the Stx 2 gene in sewage from different countries. *FEMS Microbiol Lett* 2000, 183:197-200.

133. Muniesa M, Blanco JE, De SM, Serra-Moreno R, Blanch AR, Jofre J: Diversity of stx2 converting bacteriophages induced from Shiga-toxin-producing Escherichia coli strains isolated from cattle. *Microbiology* 2004, 150:2959-2971.

134. Saldanha AJ: Java Treeview--extensible visualization of microarray data. *Bioinformatics* 2004, 20:3246-3248.

135. Khakhria R, Duck D, Lior H: Extended phage-typing scheme for Escherichia coli O157:H7. *Epidemiol Infect* 1990, 105:511-520.

136. Bolger AM, Lohse M, Usadel B: Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014, 30:2114-2120.

137. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009, 25:1754-1760.

138. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010, 20:1297-1303.

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

140. Clawson ML, Keen JE, Smith TP, Durso LM, McDaneld TG, Mandrell RE, Davis MA, Bono JL: Phylogenetic classification of Escherichia coli O157:H7 strains of human and bovine origin using a novel set of nucleotide polymorphisms. *Genome Biol* 2009, 10:R56.

141. Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, Mcvey SD, Radune D, Bergman NH, Phillippy AM: Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol* 2013, 14:R101.

142. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J: Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 2013, 10:563-569.

143. Stewart AC, Osborne B, Read TD: DIYA: a bacterial annotation pipeline for any genomics lab. *Bioinformatics* 2009, 25:962-963.

144. Luo H, Zhang CT, Gao F: Ori-Finder 2, an integrated tool to predict replication origins in the archaeal genomes. *Front Microbiol* 2014, 5:482.

145. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS: PHAST: a fast phage search tool. *Nucleic Acids Res* 2011, 39:W347-W352.

146. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: Versatile and open software for comparing large genomes. *Genome Biol* 2004, 5:R12.

147. Castanie-Cornet MP, Penfound TA, Smith D, Elliott JF, Foster JW: Control of acid resistance in Escherichia coli. *J Bacteriol* 1999, 181:3525-3535.

148. Lenski RE: Quantifying fitness and gene stability in microorganisms. *Biotechnology* 1991, 15:173-192.

149. Bochner BR, Gadzinski P, Panomitros E: Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Genome Res* 2001, 11:1246-1255.

150. Drummond AJ, Suchard MA, Xie D, Rambaut A: Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 2012, 29:1969-1973.

151. Lim JY, LA HJ, Sheng H, Forney LJ, Hovde CJ: Influence of plasmid pO157 on Escherichia coli O157:H7 Sakai biofilm formation. *Appl Environ Microbiol* 2010, 76:963-966.

152. Johnson TJ, Wannemeuhler YM, Scaccianoce JA, Johnson SJ, Nolan LK: Complete DNA sequence, comparative genomics, and prevalence of an IncHI2 plasmid occurring among extraintestinal pathogenic Escherichia coli isolates. *Antimicrob Agents Chemother* 2006, 50:3929-3933.

153. Gilmour MW, Thomson NR, Sanders M, Parkhill J, Taylor DE: The complete nucleotide sequence of the resistance plasmid R478: defining the backbone components of incompatibility group H conjugative plasmids through comparative genomics. *Plasmid* 2004, 52:182-202.

154. Feasey NA, Cain AK, Msefula CL, Pickard D, Alaerts M, Aslett M, Everett DB, Allain TJ, Dougan G, Gordon MA, Heyderman RS, Kingsley RA: Drug resistance in Salmonella enterica ser. Typhimurium bloodstream infection, Malawi. *Emerg Infect Dis* 2014, 20:1957-1959.

155. Chen YT, Lauderdale TL, Liao TL, Shiau YR, Shu HY, Wu KM, Yan JJ, Su IJ, Tsai SF: Sequencing and comparative genomic analysis of pK29, a 269-kilobase conjugative plasmid encoding CMY-8 and CTX-M-3 beta-lactamases in Klebsiella pneumoniae. *Antimicrob Agents Chemother* 2007, 51:3004-3007.

156. Li L, Liao X, Yang Y, Sun J, Li L, Liu B, Yang S, Ma J, Li X, Zhang Q, Liu Y: Spread of oqxAB in Salmonella enterica serotype Typhimurium predominantly by IncHI2 plasmids. *J Antimicrob Chemother* 2013, 68:2263-2268.

157. Kariuki S, Okoro C, Kiiru J, Njoroge S, Omuse G, Langridge G, Kingsley RA, Dougan G, Revathi G: Ceftriaxone-resistant Salmonella enterica serotype typhimurium sequence type 313 from Kenyan patients is associated with the blaCTX-M-15 gene on a novel IncHI2 plasmid. *Antimicrob Agents Chemother* 2015, 59:3133-3139.

158. Whelan KF, Sherburne RK, Taylor DE: Characterization of a region of the IncHI2 plasmid R478 which protects Escherichia coli from toxic effects specified

by components of the tellurite, phage, and colicin resistance cluster. *J Bacteriol* 1997, 179:63-71.

159. Whelan KF, Colleran E, Taylor DE: Phage inhibition, colicin resistance, and tellurite resistance are encoded by a single cluster of genes on the IncHI2 plasmid R478. *J Bacteriol* 1995, 177:5016-5027.

160. Fang L, Li X, Li L, Li S, Liao X, Sun J, Liu Y: Co-spread of metal and antibiotic resistance within ST3-IncHI2 plasmids from E. coli isolates of food-producing animals. *Sci Rep* 2016, 6:25312.

161. Richard HT, Foster JW: Acid resistance in Escherichia coli. *Adv Appl Microbiol* 2003, 52:167-186.

162. Cowley LA, Beckett SJ, Chase-Topping M, Perry N, Dallman TJ, Gally DL, Jenkins C: Analysis of whole genome sequencing for the Escherichia coli O157:H7 typing phages. *BMC Genomics* 2015, 16:271.

163. Losada L, DebRoy C, Radune D, Kim M, Sanka R, Brinkac L, Kariyawasam S, Shelton D, Fratamico PM, Kapur V, Feng PC: Whole genome sequencing of diverse Shiga toxin-producing and non-producing Escherichia coli strains reveals a variety of virulence and novel antibiotic resistance plasmids. *Plasmid* 2016, 83:8-11.

164. Ferenci T: Maintaining a healthy SPANC balance through regulatory and mutational adaptation. *Mol Microbiol* 2005, 57:1-8.

165. Corbishley A, Ahmad NI, Hughes K, Hutchings MR, McAteer SP, Connelley TK, Brown H, Gally DL, McNeilly TN: Strain-dependent cellular immune responses in cattle following Escherichia coli O157:H7 colonization. *Infect Immun* 2014, 82:5117-5131.

166. Barquist L, Mayho M, Cummins C, Cain AK, Boinett C, Page AJ, Langridge GC, Quail MA, Keane JA, Parkhill J: The TraDIS toolkit: sequencing and analysis for dense transposon mutant libraries. *Bioinformatics* 2016.

167. Storey JD, Tibshirani R: Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 2003, 100:9440-9445.

168. Tatusov RL, Koonin EV, Lipman DJ: A genomic perspective on protein families. *Science* 1997, 278:631-637.

169. Parra-Lopez C, Baer MT, Groisman EA: Molecular genetic analysis of a locus required for resistance to antimicrobial peptides in Salmonella typhimurium. *EMBO J* 1993, 12:4053-4062.

170. Yu F, Yamada H, Mizushima S: Role of lipopolysaccharide in the receptor function for bacteriophage TuIb in Escherichia coli. *J Bacteriol* 1981, 148:712-715.

171. Rossmann MG, Mesyanzhinov VV, Arisaka F, Leiman PG: The bacteriophage T4 DNA injection machine. *Curr Opin Struct Biol* 2004, 14:171-180.

172. Mizuno T, Kato M, Jo YL, Mizushima S: Interaction of OmpR, a positive regulator, with the osmoregulated ompC and ompF genes of Escherichia coli. Studies with wild-type and mutant OmpR proteins. *J Biol Chem* 1988, 263:1008-1012.

173. Igo MM, Ninfa AJ, Silhavy TJ: A bacterial environmental sensor that functions as a protein kinase and stimulates transcriptional activation. *Genes Dev* 1989, 3:598-605.

174. Williams MD, Fuchs JA, Flickinger MC: Null mutation in the stringent starvation protein of Escherichia coli disrupts lytic development of bacteriophage P1. *Gene* 1991, 109:21-30.

175. Larranaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armananzas R, Santafe G, Perez A, Robles V: Machine learning in bioinformatics. *Brief Bioinform* 2006, 7:86-112.

176. Szymczak S, Biernacka JM, Cordell HJ, Gonzalez-Recio O, Konig IR, Zhang H, Sun YV: Machine learning in genome-wide association studies. *Genet Epidemiol* 2009, 33 Suppl 1:S51-S57.

177. Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R, Henderson IR, Sperandio V, Ravel J: The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. *J Bacteriol* 2008, 190:6881-6893.

178. D'Auria G, Jimenez-Hernandez N, Peris-Bondia F, Moya A, Latorre A: Legionella pneumophila pangenome reveals strain-specific virulence factors. *BMC Genomics* 2010, 11:181.

Chapter 7: Appendix

7.1 Availability of data

Chapter 2: The raw sequencing reads have been deposited in the short read archive under project alias PRJNA252693. The assembled sequences and annotations of the typing phages have been deposited in GenBank under the following accessions;

Phage 1: KP869100

Phage 3: KP869101

Phage 4: KP869102

Phage 5: KP869103

Phage 6: KP869104

Phage 7: KP869105

Phage 8: KP869106

Phage 9: KP869107

Phage 10: KP869108

Phage 11: KP869109

Phage 12: KP869110

Phage 13: KP869111

Phage 14: KP869112

Phage 15: KP869113

Chapter 4: FASTQ sequences were deposited in the NCBI Short Read Archive under the BioProject PRJNA248042. 2. Long read FASTA files are deposited in NCBI Genbank under accessions CP015831 (644-PT8 chromosome), CP015832 (180-PT54 chromosome) and CP015833(180-PT54 plasmid).

Table 7.1 Table detailing the accessory variation of Group 2 as depicted in Figure 2.8. Presence of accessory gene in phage genome depicted by 'X' and absence by a blank box.

Group 2	Phage 3	Phage 6	Phage 7	Phage 13
PROKKA7_00005_Phage_tail_fibre_adhesin_Gp38			X	
PROKKA7_00008_hypothetical_protein			X	
PROKKA3_00008_hypothetical_protein	X	X		X
PROKKA7_00009_hypothetical_protein			X	
PROKKA7_00010_hypothetical_protein			X	
PROKKA7_00011_hypothetical_protein			X	
PROKKA3_00011_hypothetical_protein	X	X		X
PROKKA7_00015_hypothetical_protein			X	
PROKKA6_00016_hypothetical_protein	X	X		X
PROKKA3_00016_hypothetical_protein	X	X		X
PROKKA7_00017_hypothetical_protein			X	
PROKKA7_00019_hypothetical_protein			X	
PROKKA7_00020_hypothetical_protein			X	
PROKKA7_00021_hypothetical_protein			X	
PROKKA3_00021_hypothetical_protein	X	X		X
PROKKA7_00022_hypothetical_protein			X	
PROKKA7_00023_hypothetical_protein			X	
PROKKA3_00023_hypothetical_protein	X	X		X

PROKKA3_00024_hypothetical_protein	X	X		X
PROKKA7_00026_hypothetical_protein			X	
PROKKA7_00027_hypothetical_protein			X	
PROKKA3_00028_Phospho-2-dehydro-3-deoxyheptonate_aldolase_Tyr-sensitive	X	X		X
PROKKA3_00030_hypothetical_protein	X	X		X
PROKKA7_00032_hypothetical_protein			X	
PROKKA7_00033_hypothetical_protein			X	
PROKKA3_00034_Phage_GP30.8_protein	X	X		X
PROKKA3_00036_hypothetical_protein	X	X		X
PROKKA7_00037_hypothetical_protein			X	
PROKKA7_00038_hypothetical_protein			X	
PROKKA3_00038_hypothetical_protein	X	X		X
PROKKA7_00040_hypothetical_protein			X	
PROKKA3_00041_hypothetical_protein	X	X		X
PROKKA7_00042_hypothetical_protein			X	
PROKKA7_00043_hypothetical_protein			X	
PROKKA7_00044_hypothetical_protein			X	
PROKKA3_00045_hypothetical_protein	X	X		X
PROKKA7_00048_hypothetical_protein			X	
PROKKA7_00049_hypothetical_protein			X	
PROKKA7_00050_hypothetical_protein			X	
PROKKA7_00051_hypothetical_protein			X	
PROKKA7_00052_hypothetical_protein			X	
PROKKA7_00053_hypothetical_protein			X	

PROKKA7_00054_hypothetical_protein			X	
PROKKA7_00055_hypothetical_protein			X	
PROKKA7_00056_hypothetical_protein			X	
PROKKA7_00058_hypothetical_protein			X	
PROKKA7_00060_hypothetical_protein			X	
PROKKA3_00062_hypothetical_protein	X	X		X
PROKKA3_00065_hypothetical_protein	X	X		X
PROKKA7_00066_hypothetical_protein			X	
PROKKA7_00071_hypothetical_protein			X	
PROKKA7_00072_Thymidylate_synthase_1			X	
PROKKA7_00073_hypothetical_protein			X	
PROKKA7_00074_hypothetical_protein			X	
PROKKA7_00075_hypothetical_protein			X	
PROKKA7_00077_hypothetical_protein			X	
PROKKA7_00082_Phage_RNA_polymerase_binding_RpbA			X	
PROKKA7_00085_hypothetical_protein			X	
PROKKA7_00086_hypothetical_protein			X	
PROKKA7_00088_DNA_alpha-glucosyltransferase			X	
PROKKA7_00091_hypothetical_protein			X	
PROKKA7_00093_hypothetical_protein			X	
PROKKA3_00097_Nuclease_inhibitor_from_bacteriophage_T4	X	X		X
PROKKA3_00098_hypothetical_protein	X	X		
PROKKA3_00099_hypothetical_protein	X			
PROKKA7_00099_hypothetical_protein			X	

PROKKA3_00100_hypothetical_protein	X			
PROKKA3_00101_hypothetical_protein	X			
PROKKA3_00102_hypothetical_protein	X			
PROKKA3_00103_tRNA-Met(cat)	X			
PROKKA3_00104_tRNA-Arg(tct)	X			
PROKKA3_00105_hypothetical_protein	X			
PROKKA3_00106_hypothetical_protein	X			
PROKKA3_00107_hypothetical_protein	X			
PROKKA7_00107_hypothetical_protein			X	
PROKKA3_00108_hypothetical_protein	X			
PROKKA3_00109_Bacteriophage_FRD3_protein	X			
PROKKA7_00109_hypothetical_protein			X	
PROKKA3_00110_hypothetical_protein	X			
PROKKA3_00111_hypothetical_protein	X			
PROKKA7_00111_hypothetical_protein			X	
PROKKA3_00112_hypothetical_protein	X			
PROKKA7_00112_hypothetical_protein			X	
PROKKA3_00113_hypothetical_protein	X			
PROKKA3_00114_hypothetical_protein	X			
PROKKA7_00114_hypothetical_protein			X	
PROKKA3_00115_hypothetical_protein	X			
PROKKA7_00115_hypothetical_protein			X	
PROKKA3_00116_hypothetical_protein	X			
PROKKA7_00116_hypothetical_protein			X	

PROKKA7_00117_hypothetical_protein			X	
PROKKA3_00117_hypothetical_protein	X			X
PROKKA3_00118_hypothetical_protein	X			
PROKKA3_00119_hypothetical_protein	X			X
PROKKA7_00119_hypothetical_protein			X	
PROKKA3_00120_hypothetical_protein	X	X		X
PROKKA3_00121_hypothetical_protein	X	X		X
PROKKA3_00122_hypothetical_protein	X	X		X
PROKKA3_00123_hypothetical_protein	X	X		
PROKKA7_00123_hypothetical_protein			X	
PROKKA7_00125_hypothetical_protein			X	
PROKKA7_00126_hypothetical_protein			X	
PROKKA3_00126_hypothetical_protein	X	X		X
PROKKA3_00128_hypothetical_protein	X	X		
PROKKA7_00128_hypothetical_protein			X	
PROKKA7_00129_hypothetical_protein			X	
PROKKA7_00130_hypothetical_protein			X	
PROKKA7_00131_hypothetical_protein			X	
PROKKA7_00132_hypothetical_protein			X	
PROKKA7_00136_hypothetical_protein			X	
PROKKA7_00137_hypothetical_protein			X	
PROKKA7_00138_hypothetical_protein			X	
PROKKA3_00141_hypothetical_protein	X	X		X
PROKKA3_00145_hypothetical_protein	X	X		X

PROKKA3_00146_hypothetical_protein	X	X		X
PROKKA3_00148_hypothetical_protein	X	X		X
PROKKA7_00151_hypothetical_protein			X	
PROKKA3_00152_hypothetical_protein	X	X		X
PROKKA3_00154_hypothetical_protein	X	X		X
PROKKA7_00155_hypothetical_protein			X	
PROKKA3_00156_hypothetical_protein	X	X		X
PROKKA7_00156_hypothetical_protein			X	
PROKKA7_00157_hypothetical_protein			X	
PROKKA3_00157_hypothetical_protein	X	X		X
PROKKA3_00159_hypothetical_protein	X	X		X
PROKKA3_00160_hypothetical_protein	X	X		X
PROKKA7_00160_hypothetical_protein			X	
PROKKA7_00161_hypothetical_protein			X	
PROKKA3_00163_hypothetical_protein	X	X		X
PROKKA7_00164_tRNA-Arg(tct)			X	
PROKKA3_00165_hypothetical_protein	X	X		X
PROKKA7_00165_tRNA-His(gtg)			X	
PROKKA3_00166_hypothetical_protein	X	X		X
PROKKA7_00166_tRNA-Asn(gtt)			X	
PROKKA7_00167_tRNA-Tyr(gta)			X	
PROKKA7_00168_tRNA-Met(cat)			X	
PROKKA7_00169_tRNA-Thr(tgt)			X	
PROKKA7_00170_tRNA-Ser(tga)			X	

PROKKA7_00171_tRNA-Pro(tgg)			X	
PROKKA7_00172_tRNA-Gly(tcc)			X	
PROKKA7_00173_tRNA-Leu(taa)			X	
PROKKA3_00173_hypothetical_protein	X	X		X
PROKKA7_00174_tRNA-Gln(ttg)			X	
PROKKA3_00174_hypothetical_protein	X	X		X
PROKKA7_00176_hypothetical_protein			X	
PROKKA3_00178_hypothetical_protein	X	X		X
PROKKA7_00178_hypothetical_protein			X	
PROKKA3_00181_hypothetical_protein	X	X		X
PROKKA3_00183_hypothetical_protein	X	X		X
PROKKA3_00186_hypothetical_protein	X	X		X
PROKKA3_00188_hypothetical_protein	X	X		X
PROKKA3_00189_hypothetical_protein	X	X		X
PROKKA3_00192_Phage_RNA_polymerase_binding_RpbA	X	X		X
PROKKA3_00198_Arabinose_5-phosphate_isomerase_KpsF	X	X		X
PROKKA3_00199_hypothetical_protein	X	X		X
PROKKA3_00200_CTP:phosphocholine_cytidylyltransferase_involved_in_choline_phosphorylation_for_cell_surface_LPS_epitopes	X	X		X
PROKKA3_00201_capsule_biosynthesis_phosphatase	X	X		X
PROKKA3_00202_Collagenase	X	X		X
PROKKA3_00203_hypothetical_protein	X	X		X
PROKKA3_00204_Thymidylate_synthase	X	X		X
PROKKA3_00205_hypothetical_protein	X	X		X
PROKKA3_00206_hypothetical_protein	X	X		X

PROKKA3_00210_hypothetical_protein	X	X		X
PROKKA3_00211_hypothetical_protein	X	X		X
PROKKA7_00212_hypothetical_protein			X	
PROKKA3_00213_hypothetical_protein	X	X		X
PROKKA3_00214_hypothetical_protein	X	X		X
PROKKA3_00217_hypothetical_protein	X	X		X
PROKKA3_00220_hypothetical_protein	X	X		X
PROKKA3_00222_hypothetical_protein	X	X		X
PROKKA3_00223_hypothetical_protein	X	X		X
PROKKA3_00227_hypothetical_protein	X	X		X
PROKKA3_00228_hypothetical_protein	X	X		X
PROKKA3_00229_hypothetical_protein	X	X		X
PROKKA7_00229_hypothetical_protein			X	
PROKKA7_00230_ADP-ribosyltransferase_exoenzyme			X	
PROKKA3_00230_hypothetical_protein	X	X		X
PROKKA3_00231_hypothetical_protein	X	X		X
PROKKA3_00232_hypothetical_protein	X	X		X
PROKKA7_00234_hypothetical_protein			X	
PROKKA7_00237_hypothetical_protein			X	
PROKKA3_00238_hypothetical_protein	X	X		X
PROKKA3_00239_hypothetical_protein	X	X		X
PROKKA7_00239_hypothetical_protein			X	
PROKKA3_00240_hypothetical_protein	X	X		X
PROKKA7_00241_Phage_GP30.8_protein			X	

PROKKA3_00244_hypothetical_protein	X	X		
PROKKA7_00245_hypothetical_protein			X	
PROKKA7_00246_hypothetical_protein			X	
PROKKA3_00246_hypothetical_protein	X	X		X
PROKKA7_00248_hypothetical_protein			X	
PROKKA3_00250_hypothetical_protein	X	X		X
PROKKA7_00251_hypothetical_protein			X	
PROKKA7_00252_hypothetical_protein			X	
PROKKA3_00253_hypothetical_protein	X	X		X
PROKKA3_00254_hypothetical_protein	X	X		X
PROKKA3_00256_hypothetical_protein	X	X		X
PROKKA7_00258_hypothetical_protein			X	
PROKKA3_00261_hypothetical_protein	X	X		X
PROKKA3_00262_hypothetical_protein	X	X		X
PROKKA3_00263_hypothetical_protein	X	X		X
PROKKA7_00263_hypothetical_protein			X	
PROKKA3_00264_hypothetical_protein	X	X		X
PROKKA7_00266_hypothetical_protein			X	
PROKKA3_00267_Caudovirales_tail_fibre_assembly_protein	X	X		X
PROKKA7_00268_hypothetical_protein			X	
PROKKA3_00268_Phage_Tail_Collar_Domain_protein	X	X		
PROKKA7_00269_hypothetical_protein			X	
PROKKA7_00271_hypothetical_protein			X	
PROKKA7_00273_Bacteriophage_FRD3_protein			X	

PROKKA7_00279_hypothetical_protein			X	
PROKKA3_00277_Bacteriophage_replication_gene_A_protein_(GPA)	X	X		X
PROKKA7_00285_hypothetical_protein			X	

Table 7.2. Table detailing the unique chromosomal genes found in the PT8 and PT54 genes with their annotation and chromosome location

PT8 Unique genes	PT8 Genome location	PT54 Unique genes	PT54 Genome location
644_00713_hypothetical_protein	722431-722838	180_00068_ilvB_operon_leader_peptide	75339-75428
644_02073_hypothetical_protein	2102929-2103033	180_00338_hypothetical_protein	356642-356917
644_02384_hypothetical_protein	2375234-2375443	180_02292_Phage_tail_sheath_protein	2286868-2288052
644_02879_hypothetical_protein	2841585-2841818	180_02293_Phage_tail_tube_protein_FII	2288052-2288564
644_02880_hypothetical_protein	2841796-2842203	180_02294_Phage_tail_protein_E	2288619-2288984
644_02883_transcriptional_repressor_DicA	2843080-2843487	180_02295_Phage-related_minor_tail_protein	2289135-2291225
644_02889_hypothetical_protein	2847420-2848178	180_02296_hypothetical_protein	2291237-2291938
644_02892_hypothetical_protein	2849217-2849495	180_02297_Phage_P2_GpU	2291951-2292439
644_02895_hypothetical_protein	2850924-2851454	180_02298_Serine_acetyltransferase	2292596-2293168
644_03322_hypothetical_protein	3265153-3265803	180_02302_Plasmid_stability_protein	2294834-2295148
644_03323_IpaB/EvcA_family_protein	3267310-3267900	180_02303_Plasmid_segregation_protein_ParM	2295153-2296112
644_03324_hypothetical_protein	3268084-3268731	180_02304_Bacteriophage_replication_gene_A_protein_(GPA)	2296189-2299011
644_03325_hypothetical_protein	3269486-3269755	180_02305_hypothetical_protein	2299018-2299383
644_03329_Transposase,_Mutator_family	3272657-3273004	180_02306_hypothetical_protein	2299456-2299686
644_03331_Effector_protein_NleF	3273034-3273291	180_02307_hypothetical_protein	2300009-2300308

644_03333_Tyrosine_recombinase_XerD	3275358-3276050	180_02308_hypothetical_protein	2300305-2300571
644_03334_hypothetical_protein	3276711-3278036	180_02309_hypothetical_protein	2300568-2300771
644_03347_hypothetical_protein	3291283-3291399	180_02310_hypothetical_protein	2300795-2301211
644_03371_hypothetical_protein	3308835-3309365	180_02311_hypothetical_protein	2301304-2301417
644_03374_hypothetical_protein	3310794-3311072	180_02312_hypothetical_protein	2301414-2301656
644_03377_hypothetical_protein	3312111-3312869	180_02313_hypothetical_protein	2301668-2301946
644_03383_transcriptional_repressor_DicA	3316802-3317209	180_02314_hypothetical_protein	2301957-2302307
644_03386_hypothetical_protein	3318086-3318493	180_02315_hypothetical_protein	2302329-2302532
644_03387_hypothetical_protein	3318471-3318704	180_02316_Helix-turn-helix_domain_protein	2302831-2303235
644_03792_hypothetical_protein	3639232-3639387	180_02317_flagella_biosynthesis_regulator	2303251-2303901
644_04784_hypothetical_protein	4717144-4717524	180_02318_hypothetical_protein	2303931-2304278
644_04785_hypothetical_protein	4717566-4718660	180_02319_Tyrosine_recombinase_XerC	2304284-2305285
644_04786_hypothetical_protein	4718614-4718826	180_02407_hypothetical_protein	2390892-2391014
644_04787_hypothetical_protein	4719008-4719424	180_03074_hypothetical_protein	3048749-3048847
644_04788_Lactate_utilization_protein_C	4719919-4720614	180_03076_hypothetical_protein	3049616-3050173
644_04789_Lactate_utilization_protein_B	4720607-4722034		
644_04790_Lactate_utilization_protein_A	4722045-4722764		
644_04791_Virulence_regulon_transcriptional_activator_VirF	4723292-4724146		
644_04792_Mercuric_reductase	4724372-4725697		
644_04793_Inner_membrane_protein_YkgB	4726054-4726647		
644_04794_putative_oxidoreductase_YtbE	4726807-4727676		

644_04795_Right_origin-binding_protein	4727925-4728782		
644_04796_Invasin	4728903-4733156		
644_04797_Glyoxal_reductase	4733722-4734297		
644_04798_hypothetical_protein	4734490-4734846		
644_04799_HTH-type_transcriptional_regulator_DmlR	4735134-4736060		
644_04800_Alpha/beta_hydrolase_family_protein	4736217-4737137		
644_04801_NADH_oxidase	4737372-4738514		
644_04805_50S_ribosomal_protein_L31_type_B	4741899-4742165		
644_04806_50S_ribosomal_protein_L36_2	4742165-4742305		
644_04807_hypothetical_protein	4742375-4742566		
644_04808_HTH-type_transcriptional_regulator_MatA	4743391-4743933		
644_04809_hypothetical_protein	4744008-4744595		
644_04810_hypothetical_protein	4744653-4745321		
644_04811_hypothetical_protein	4745347-4747872		
644_04812_hypothetical_protein	4747862-4748442		
644_04813_hypothetical_protein	4749474-4750184		
644_04814_Inner_membrane_protein_YagU	4751074-4751688		
644_04815_Carbon_monoxide_dehydrogenase_small_chain	4752106-4752795		
644_04816_4-hydroxybenzoyl-CoA_reductase_subunit_beta	4752792-4753748		
644_04817_Xanthine_dehydrogenase_molybdenum-binding_subunit	4753745-4755943		
644_04818_XdhC_and_CoxI_family_protein	4755953-4756909		

644_04819_Purine_ribonucleoside_efflux_pump_NepI	4757088-4758215		
644_04820_Alpha/beta_hydrolase_family_protein	4758357-4759415		
644_04821_HTH-type_transcriptional_regulator_DmlR	4759661-4760563		
644_04822_hypothetical_protein	4761266-4761544		
644_04823_putative_oxidoreductase	4761711-4762433		
644_04824_HTH-type_transcriptional_regulator_DmlR	4762532-4763431		
644_04825_hypothetical_protein	4764107-4765063		
644_04826_DNA_primase_TraC	4765196-4767529		
644_04827_hypothetical_protein	4767543-4767866		
644_04828_hypothetical_protein	4767866-4768087		
644_04829_Ash_protein_family_protein	47680084-4768641		
644_04830_Prophage_CP4-57_regulatory_protein_(AlpA)	4768638-4768898		
644_04831_Phage_polarity_suppression_protein_(Psu)	4769832-4770584		
644_04832_Phage_polarity_suppression_protein_(Psu)	4770581-4771132		
644_04833_DNA-binding_transcriptional_regulator	4771138-4771410		
644_04834_hypothetical_protein	4771820-4772836		
644_04835_hypothetical_protein	4772386-4772976		
644_04836_hypothetical_protein	4773007-4773639		
644_04837_hypothetical_protein	4773632-4774090		
644_04838_hypothetical_protein	4774090-4774707		
644_04839_hypothetical_protein	4774680-4775096		

644_04840_Putative_prophage_CPS-53_integrase	4775100-4776281		
644_04841_IS2_transposase_TnpB	4776849-4777202		
644_04842_HTH-type_transcriptional_regulator_YdeO	4777244-4777987		
644_04843_hypothetical_protein	4778811-4779584		

Table 7.3 Table showing the 114 genes highlighted by the TraDIS selections as being involved in bacteriophage infection. The table shows the gene locus name, the COG group that it has been assigned to, the annotation of that gene and the Log fold change compared to the control in number of insertions that was observed after bacteriophage selection.

Gene locus	**COG Group**	**Annotation**	**LogFC vs the control**
Ecoli9000q_16260	CELL WALL/MEMBRANE/ENVELOPE BIOGENESIS	Outer membrane protein C	9.948519808
Ecoli9000q_29070	SIGNAL TRANSDUCTION MECHANISMS	Osmolarity sensor protein EnvZ	9.812073088
Ecoli9000q_29080	SIGNAL TRANSDUCTION MECHANISMS/TRANSCRIPTION	Transcriptional regulatory protein OmpR	9.413818163

Ecoli9000q_26980	TRANSLATION, RIBOSOMAL STRUCTURE AND BIOGENESIS	hypothetical protein	8.479335941
Ecoli9000q_3240	DEFENSE MECHANISMS	Peptide transport system permease protein sapC	8.379641001
Ecoli9000q_28120	INORGANIC ION TRANSPORT AND METABOLISM	Trk system potassium uptake protein TrkA	7.975556825
Ecoli9000q_34640	INORGANIC ION TRANSPORT AND METABOLISM	Potassium uptake protein, TrkH	7.504031269
Ecoli9000q_3230	CARBOHYDRATE TRANSPORT AND METABOLISM	Peptide transport system ATP-binding protein SapD	7.324206894
Ecoli9000q_19710	TRANSCRIPTION/SIGNAL TRANSDUCTION MECHANISMS	Hydrogenase-4 transcriptional activator	7.305550369
Ecoli9000q_35390	CARBOHYDRATE TRANSPORT AND METABOLISM	6-phosphofructokinase	7.192950724

Ecoli9000q_3220	CELL WALL/MEMBRANE/ENVELOPE BIOGENESIS	Peptide transport system ATP-binding protein SapF	6.923231391
Ecoli9000q_3260	DEFENSE MECHANISMS	Peptide transport periplasmic protein sapA	6.620230847
Ecoli9000q_36030	TRANSLATION, RIBOSOMAL STRUCTURE AND BIOGENESIS	Elongation factor Tu	6.174319841
Ecoli9000q_10280	TRANSCRIPTION	Lipid A biosynthesis (KDO)2-(lauroyl)-lipid IVA acyltransferase	6.172719132
Ecoli9000q_27370	CARBOHYDRATE TRANSPORT AND METABOLISM/SIGNAL TRANSDUCTION MECHANISMS	Nitrogen regulatory protein	5.853102337
Ecoli9000q_34500	COENZYME TRANSPORT AND METABOLISM	Ubiquinone/menaquinone biosynthesis methyltransferase ubiE	5.832565746

Ecoli9000q_54460	CELL MOTILITY	Ribonuclease E	5.542857839
Ecoli9000q_20540	COENZYME TRANSPORT AND METABOLISM	Pyridoxine 5'-phosphate synthase	5.534946763
Ecoli9000q_3250	DEFENSE MECHANISMS	Peptide transport system permease protein sapB	5.520564259
Ecoli9000q_21070	POSTRANSLATIONAL MODIFICATION, PROTEIN TURNOVER, CHAPERONES	SsrA-binding protein	5.484968276
Ecoli9000q_14240	TRANSCRIPTION / CARBOHYDRATE TRANSPORT AND METABOLISM	Galactitol permease IIC component	5.464163309
Ecoli9000q_47320	GENERAL FUNCTION PREDICTION ONLY	Lipoyl synthase	5.051970875
Ecoli9000q_50850	AMINO ACID TRANSPORT AND METABOLISM	Cytidylate kinase	5.037706786
Ecoli9000q_t350	NO COG HIT		5.019387592
Ecoli9000q_24250	COENZYME TRANSPORT AND METABOLISM/ENERGY	Protein visC	4.919457267

	PRODUCTION AND CONVERSION		
Ecoli9000q_28660	SIGNAL TRANSDUCTION MECHANISMS	Catabolite gene activator	4.899718743
Ecoli9000q_8120	ENERGY PRODUCTION AND CONVERSION	Ribonuclease T	4.658085247
Ecoli9000q_3150	SIGNAL TRANSDUCTION MECHANISMS	protein yciW	4.552476946
Ecoli9000q_34580	COENZYME TRANSPORT AND METABOLISM	3-octaprenyl-4-hydroxybenzoate carboxy-lyase	4.418526722
Ecoli9000q_17060	GENERAL FUNCTION PREDICTION ONLY	Phosphate acetyltransferase	4.394529392
Ecoli9000q_24260	COENZYME TRANSPORT AND METABOLISM/ENERGY PRODUCTION AND CONVERSION	2-octaprenyl-6-methoxyphenol hydroxylase	4.392847702
Ecoli9000q_28080	REPLICATION, RECOMBINATION AND REPAIR	Protein smf	4.203530306
Ecoli9000q_54090	LIPID TRANSPORT AND METABOLISM	protein YmdC	4.201746198

Ecoli9000q_27430	SIGNAL TRANSDUCTION MECHANISMS	Aerobic respiration control sensor protein ArcB	4.016808495
Ecoli9000q_36020	COENZYME TRANSPORT AND METABOLISM	Pantothenate kinase	3.916358073
Ecoli9000q_34510	COENZYME TRANSPORT AND METABOLISM	protein yigP	3.868346125
Ecoli9000q_13490	MOBILOME: PROPHAGES, TRANSPOSONS	Exodeoxyribonuclease I	3.849851565
Ecoli9000q_24220	AMINO ACID TRANSPORT AND METABOLISM	Glycine cleavage system H protein	3.788714647
Ecoli9000q_16960	ENERGY PRODUCTION AND CONVERSION	reductase	3.785724884
Ecoli9000q_42320	CELL WALL/MEMBRANE/ENVELOPE BIOGENESIS / POSTRANSLATIONAL MODIFICATION, PROTEIN TURNOVER, CHAPERONES	Chaperone protein skp	3.643697255
Ecoli9000q_41690	ENERGY PRODUCTION AND CONVERSION	Aconitate hydratase 2	3.59398253

Ecoli9000q_19900	SIGNAL TRANSDUCTION MECHANISMS	Inosine-5'-monophosphate dehydrogenase	3.58964449
Ecoli9000q_1170	GENERAL FUNCTION PREDICTION ONLY	Protease 7	3.546143167
Ecoli9000q_24070	TRANSLATION, RIBOSOMAL STRUCTURE AND BIOGENESIS	Lysyl-tRNA synthetase	3.497873184
Ecoli9000q_31710	INTRACELLULAR TRAFFICKING, SECRETION, AND VESICULAR TRANSPORT	SecB protein	3.468424177
Ecoli9000q_54110	LIPID TRANSPORT AND METABOLISM	Glucans biosynthesis protein G	3.465076834
Ecoli9000q_54720	ENERGY PRODUCTION AND CONVERSION	HTH-type transcriptional regulator ycfQ	3.451890337
Ecoli9000q_29660	AMINO ACID TRANSPORT AND METABOLISM	High-affinity branched-chain amino acid transport ATP-	3.34686467

		binding protein LivG	
Ecoli9000q_36230	FUNCTION UNKNOWN	protein yjaG	3.314032794
Ecoli9000q_41120	TRANSCRIPTION/ REPLICATION, RECOMBINATION AND REPAIR	RNA polymerase-associated protein rapA	3.309043081
Ecoli9000q_41070	POSTRANSLATIONAL MODIFICATION, PROTEIN TURNOVER, CHAPERONES	Chaperone surA	3.236827539
Ecoli9000q_1790	INORGANIC ION TRANSPORT AND METABOLISM	transport protein	3.179788013
Ecoli9000q_10730	CARBOHYDRATE TRANSPORT AND METABOLISM	Ferritin-like protein 2	3.175587189
Ecoli9000q_41650	ENERGY PRODUCTION AND CONVERSION	Pyruvate dehydrogenase E1 component	3.143582779

Ecoli9000q_40500	SIGNAL TRANSDUCTION MECHANISMS/TRANSCRIPTION	Aerobic respiration control protein ArcA	3.098701665
Ecoli9000q_45750	NO COG HIT	Hemolysin expression-modulating protein Hha	3.096482664
Ecoli9000q_14440	CELL MOTILITY/EXTRACELLULAR STRUCTURE	outer membrane usher protein yehB	3.096437589
Ecoli9000q_590	SIGNAL TRANSDUCTION MECHANISMS/TRANSCRIPTION	High frequency lysogenization protein HflD	3.055797576
Ecoli9000q_47830	INORGANIC ION TRANSPORT AND METABOLISM	Ferric uptake regulation protein	3.043047177
Ecoli9000q_13630	AMINO ACID TRANSPORT AND METABOLISM	Chain length determinant protein	2.979668465
Ecoli9000q_20830	TRANSLATION, RIBOSOMAL STRUCTURE AND BIOGENESIS	Pseudouridine synthase	2.953027476

Ecoli9000q_26570	NO COG HIT	Toxin YhaV	2.943455374
Ecoli9000q_31870	COENZYME TRANSPORT AND METABOLISM/SIGNAL TRANSDUCTION MECHANISMS	Lipopolysaccharide core biosynthesis protein rfaY	2.925611269
Ecoli9000q_4420	AMINO ACID TRANSPORT AND METABOLISM	N-acetyltransferase YncA	2.857491414
Ecoli9000q_33430	INORGANIC ION TRANSPORT AND METABOLISM	hypothetical protein	2.83383681
Ecoli9000q_9660	NO COG HIT	hypothetical protein	2.813303769
Ecoli9000q_26220	CARBOHYDRATE TRANSPORT AND METABOLISM	Uronate isomerase	2.802230473
Ecoli9000q_35930	TRANSCRIPTION	HTH-type transcriptional repressor fabR	2.78749035
Ecoli9000q_33450	INORGANIC ION TRANSPORT AND METABOLISM	Phosphate transport system permease protein pstC	2.741594751

Ecoli9000q_28110	TRANSLATION, RIBOSOMAL STRUCTURE AND BIOGENESIS	Ribosomal RNA small subunit methyltransferase B	2.735740471
Ecoli9000q_45330	COENZYME TRANSPORT AND METABOLISM/LIPID TRANSPORT AND METABOLISM	Exodeoxyribonuclease 7 small subunit	2.692729739
Ecoli9000q_36870	NO COG HIT	Eae protein	2.653334792
Ecoli9000q_19260	AMINO ACID TRANSPORT AND METABOLISM	Ethanolamine utilization protein EutL	2.645832848
Ecoli9000q_3140	TRANSCRIPTION	Exoribonuclease 2	2.608548239
Ecoli9000q_54120	CELL WALL/MEMBRANE/ENVELOPE BIOGENESIS	Glucans biosynthesis glucosyltransferase H	2.551367208
Ecoli9000q_33400	NO COG HIT	hypothetical protein	2.534602918

Ecoli9000q_38270	CELL WALL/MEMBRANE/ENVELOPE BIOGENESIS	N-acetylmuramoyl-L-alanine amidase AmiB	2.531975021
Ecoli9000q_34250	REPLICATION, RECOMBINATION AND REPAIR	site-specific tyrosine recombinase XerC	2.531000412
Ecoli9000q_54490	TRANSLATION, RIBOSOMAL STRUCTURE BIOGENESIS	protein yceD	2.516521899
Ecoli9000q_1500	INORGANIC ION TRANSPORT AND METABOLISM	Orf2	2.485792275
Ecoli9000q_34530	INTRACELLULAR TRAFFICKING, SECRETION, AND VESICULAR TRANSPORT	hypothetical protein	2.471861812
Ecoli9000q_52250	CELL WALL/MEMBRANE/ENVELOPE BIOGENESIS	lipoprotein gfcB	2.456749205
Ecoli9000q_20320	CARBOHYDRATE TRANSPORT AND METABOLISM	protein yphB	2.433031518

Ecoli9000q_46180	POSTRANSLATIONAL MODIFICATION, PROTEIN TURNOVER, CHAPERONES	ABC transporter ATP-binding protein YbbL	2.428810586
Ecoli9000q_24940	REPLICATION, RECOMBINATION AND REPAIR/MOBILOME: PROPHAGES, TRANSPOSONS	hypothetical protein	2.381542204
Ecoli9000q_54700	CARBOHYDRATE TRANSPORT AND METABOLISM	NADH dehydrogenase	2.373597462
Ecoli9000q_33520	CELL MOTILITY	LpfA	2.371276382
Ecoli9000q_50640	ENERGY PRODUCTION AND CONVERSION/POSTRANSLATIONAL MODIFICATION, PROTEIN TURNOVER, CHAPERONES	Thioredoxin reductase	2.361072232
Ecoli9000q_19110	INORGANIC ION TRANSPORT AND METABOLISM	Sulfate transport system permease protein CysT	2.33178553

Ecoli9000q_34550	INTRACELLULAR TRAFFICKING, SECRETION, AND VESICULAR TRANSPORT	Sec-independent protein translocase protein TatC	2.307195545
Ecoli9000q_52220	CELL WALL/MEMBRANE/ENVELOPE BIOGENESIS	Polysaccharide export protein	2.303139092
Ecoli9000q_14020	MOBILOME: PROPHAGES, TRANSPOSONS	Protein AsmA	2.300268445
Ecoli9000q_29470	REPLICATION, RECOMBINATION AND REPAIR	ATP-dependent DNA helicase, RecQ	2.286268896
Ecoli9000q_38600	TRANSLATION, RIBOSOMAL STRUCTURE AND BIOGENESIS	hypothetical protein	2.259280167
Ecoli9000q_7610	ENERGY PRODUCTION AND CONVERSION	Protein ydgH	2.225616921
Ecoli9000q_10090	SIGNAL TRANSDUCTION MECHANISMS	protein yebW	2.219130252
Ecoli9000q_39720	TRANSCRIPTION	HTH-type transcriptional regulator QseD	2.213357336

Ecoli9000q_52470	NUCLEOTIDE TRANSPORT AND METABOLISM	reductase	2.208746184
Ecoli9000q_35230	CARBOHYDRATE TRANSPORT AND METABOLISM	Rhamnulose-1-phosphate aldolase	2.203717425
Ecoli9000q_27780	SIGNAL TRANSDUCTION MECHANISMS	RNase E specificity factor CsrD	2.195074895
Ecoli9000q_31910	CELL WALL/MEMBRANE/ENVELOPE BIOGENESIS	Lipopolysaccharide core heptosyltransferase rfaQ	2.185876908
Ecoli9000q_42590	SECONDARY METABOLITES BIOSYNTHESIS, TRANSPORT AND CATABOLISM	hypothetical protein	2.143596558
Ecoli9000q_49490	REPLICATION, RECOMBINATION AND REPAIR	ATP-dependent helicase dinG	2.134458526
Ecoli9000q_36740	TRANSCRIPTION/SIGNAL TRANSDUCTION MECHANISMS	LexA repressor	2.126132397

Ecoli9000q_38280	REPLICATION, RECOMBINATION AND REPAIR	DNA mismatch repair protein mutL	2.124632142
Ecoli9000q_29630	FUNCTION UNKNOWN	Tetratricopeptide TPR_2 repeat protein	2.120902593
Ecoli9000q_30030	LIPID TRANSPORT AND METABOLISM/SECONDARY METABOLITES BIOSYNTHESIS, TRANSPORT AND CATABOLISM	FabF	2.106471489
Ecoli9000q_52210	SIGNAL TRANSDUCTION MECHANISMS	Low molecular weight protein-tyrosine-phosphatase etp	2.093210765
Ecoli9000q_27380	SIGNAL TRANSDUCTION MECHANISMS	hypothetical protein	2.08572709
Ecoli9000q_34190	NUCLEOTIDE TRANSPORT AND METABOLISM	Adenylate cyclase	2.06611786
Ecoli9000q_19690	ENERGY PRODUCTION AND CONVERSION	Hydrogenase-4 component I	2.061788739

Ecoli9000q_1950	SIGNAL TRANSDUCTION MECHANISMS/TRANSCRIPTION	hypothetical protein	2.060994087
Ecoli9000q_860	MOBILOME: PROPHAGES, TRANSPOSONS	Portal protein B	2.057171473
Ecoli9000q_26270	FUNCTION UNKNOWN	Protein yqjC	2.035797293
Ecoli9000q_14530	TRANSCRIPTION	molybdate metabolism regulator	2.001034647

Table 7.4 Table showing the 44 genes highlighted by the TraDIS selections to be involved in bacteriophage resistance. The table shows the gene locus name, the COG group that it has been assigned to, the annotation of that gene and the Log fold change compared to the control in number of insertions that was observed after bacteriophage selection.

Gene Locus	**COG Group**	**Annotation**	**LogFC vs the control**
Ecoli9000q_27560	POSTRANSLATIONAL MODIFICATION, PROTEIN TURNOVER, CHAPERONES	Stringent starvation protein A	-9.544824

Ecoli9000q_48600	CARBOHYDRATE TRANSPORT AND METABOLISM	UDP-galactose 4-epimerase	-8.2397784
Ecoli9000q_13690	CARBOHYDRATE TRANSPORT AND METABOLISM	GDP-L-fucose synthetase	-8.1276917
Ecoli9000q_13710	CELL WALL/MEMBRANE/ENVELOPE BIOGENESIS	Glycosyl transferase	-7.6186096
Ecoli9000q_13670	GENERAL FUNCTION PREDICTION ONLY	Mannose-1-phosphate guanylyltransferase 2	-6.9297282
Ecoli9000q_13780	CELL WALL/MEMBRANE/ENVELOPE BIOGENESIS	UDP-N-acetylglucosamine 4-epimerase	-6.5629502
Ecoli9000q_13700	NUCLEOTIDE TRANSPORT AND METABOLISM	GDP-mannose 4,6-dehydratase	-5.9619472
Ecoli9000q_31840	CELL WALL/MEMBRANE/ENVELOPE BIOGENESIS	O-antigen ligase	-5.9264937

Ecoli9000q_13760	CELL WALL/MEMBRANE/ENVELOPE BIOGENESIS	Glycosyl transferase	-5.6220736
Ecoli9000q_38970	CARBOHYDRATE TRANSPORT AND METABOLISM	Trehalose-6-phosphate hydrolase	-5.5770472
Ecoli9000q_24770	GENERAL FUNCTION PREDICTION ONLY	hypothetical protein	-5.4304707
Ecoli9000q_13680	CARBOHYDRATE TRANSPORT AND METABOLISM	hydrolase	-5.2404962
Ecoli9000q_36540	CARBOHYDRATE TRANSPORT AND METABOLISM	Glucose-6-phosphate isomerase	-5.1642333
Ecoli9000q_33990	CELL WALL/MEMBRANE/ENVELOPE BIOGENESIS	Undecaprenyl-phosphate alpha-N-acetylglucosaminyl 1-phosphate transferase	-4.9925453
Ecoli9000q_13660	CARBOHYDRATE TRANSPORT AND METABOLISM	Phosphomannomutase	-4.7959857

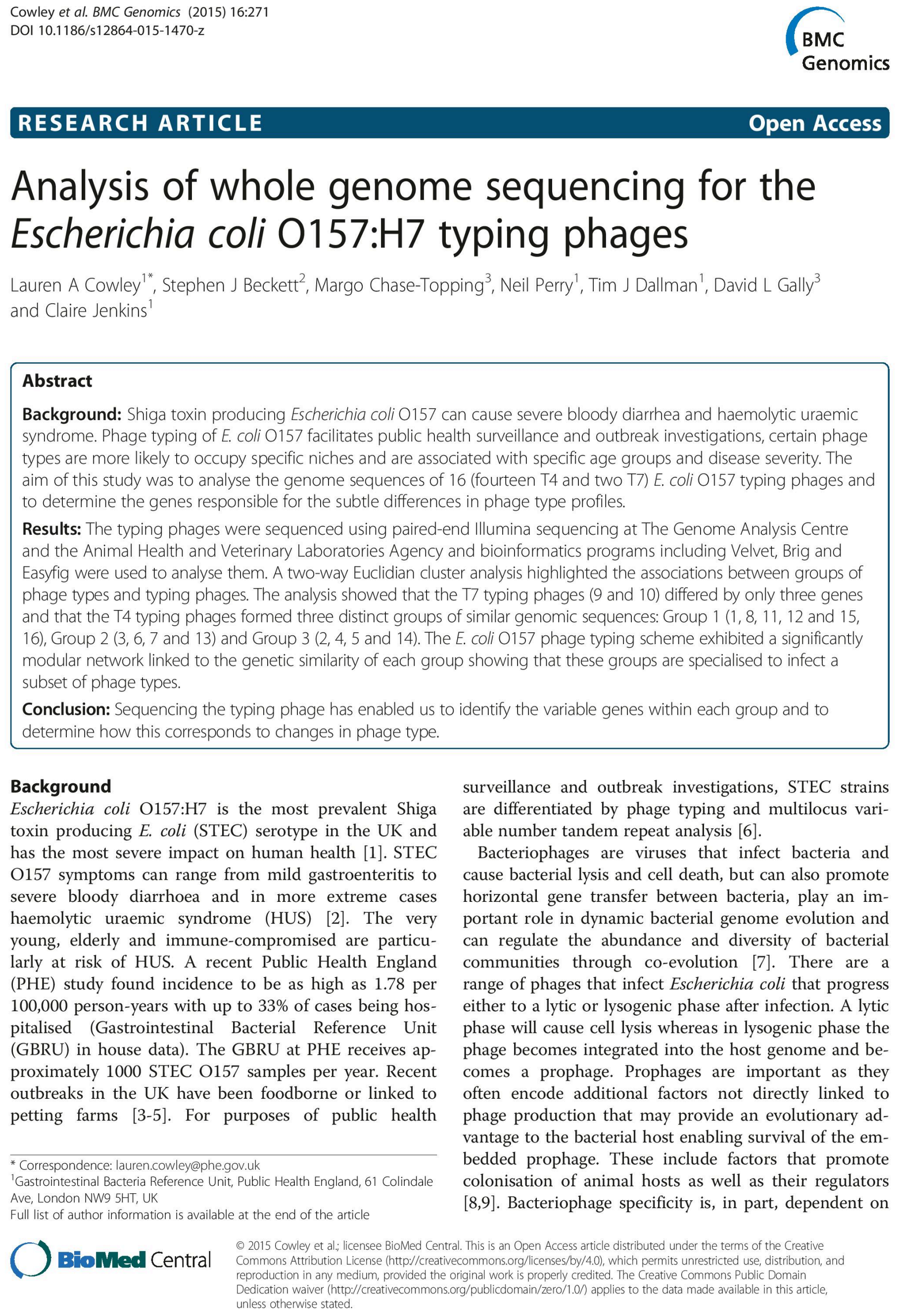
Ecoli9000q_13750	CELL WALL/MEMBRANE/ENVELOPE BIOGENESIS	O antigen polymerase Wzy	-4.7737894
Ecoli9000q_27870	TRANSLATION, RIBOSOMAL STRUCTURE AND BIOGENESIS	tRNA-dihydrouridine synthase B	-4.5789061
Ecoli9000q_31850	CELL WALL/MEMBRANE/ENVELOPE BIOGENESIS	Lipopolysaccharide 1,2-N-acetylglucosaminetransferase	-4.4548486
Ecoli9000q_8610	ENERGY PRODUCTION AND CONVERSION	Ferredoxin-like protein ydiT	-4.1433984
Ecoli9000q_31860	CELL WALL/MEMBRANE/ENVELOPE BIOGENESIS	UDP-glucose:(Galactosyl) LPS alpha1,2-glucosyltransferase WaaJ	-3.7463749
Ecoli9000q_30380	ENERGY PRODUCTION AND CONVERSION	Glutathione-disulfide reductase	-3.6298878
Ecoli9000q_34070	CELL WALL/MEMBRANE/ENVELOPE BIOGENESIS	Protein wzxE	-3.5227328

Ecoli9000q_13720	CELL WALL/MEMBRANE/ENVELOPE BIOGENESIS	Perosamine synthetase Per	-3.486645
Ecoli9000q_27140	TRANSCRIPTION	Transcription elongation factor greA	-3.3627566
Ecoli9000q_31590	CARBOHYDRATE TRANSPORT AND METABOLISM	Mannitol-1-phosphate 5-dehydrogenase	-3.2612266
Ecoli9000q_17390	TRANSLATION, RIBOSOMAL STRUCTURE AND BIOGENESIS	50S ribosomal protein L3 glutamine methyltransferase	-3.1960353
Ecoli9000q_13770	CELL WALL/MEMBRANE/ENVELOPE BIOGENESIS	UDP-glucose pyrophosphorylase	-3.0742415
Ecoli9000q_6710	NO COG HIT	Exodeoxyribonuclease VIII from bacteriophage origin	-2.9339729
Ecoli9000q_48970	NO COG HIT	LF82 chromosome, complete sequence	-2.8687045

Ecoli9000q_38290	TRANSLATION, RIBOSOMAL STRUCTURE AND BIOGENESIS	tRNA dimethylallyltransferase	-2.8565052
Ecoli9000q_15240	NO COG HIT	Conserved predicted protein	-2.8528854
Ecoli9000q_54750	REPLICATION, RECOMBINATION AND REPAIR/TRANSCRIPTION	Transcription-repair-coupling factor	-2.820876
Ecoli9000q_25700	CELL WALL/MEMBRANE/ENVELOPE BIOGENESIS	Outer membrane protein tolC	-2.7309239
Ecoli9000q_10040	TRANSCRIPTION	Protein yebR	-2.6647867
Ecoli9000q_45220	DEFENSE MECHANISMS	lipoprotein yajI	-2.5525341
Ecoli9000q_12210	DEFENSE MECHANISMS	hypothetical protein	-2.4230068
Ecoli9000q_40650	POSTTRANSLATIONAL MODIFICATION,	Chaperone protein DnaJ	-2.4227901

	PROTEIN TURNOVER, CHAPERONES		
Ecoli9000q_53350	GENERAL FUNCTION PREDICTION ONLY	L0014-like protein	-2.1913151
Ecoli9000q_6540	REPLICATION, RECOMBINATION AND REPAIR	hypothetical protein	-2.1265513
Ecoli9000q_49690	TRANSCRIPTION	L,D-transpeptidase YbiS	-2.095896
Ecoli9000q_t190	NO COG HIT		-2.0932497
Ecoli9000q_47740	NUCLEOTIDE TRANSPORT AND METABOLISM	N-acetylglucosamine-6-phosphate deacetylase	-2.0779462
Ecoli9000q_6510	REPLICATION, RECOMBINATION AND REPAIR	Antitermination protein	-2.0665247
Ecoli9000q_22540	NO COG HIT	hypothetical protein	-2.0457362

7.2 Published material from thesis



Cowley *et al. BMC Genomics* (2015) 16:271
DOI 10.1186/s12864-015-1470-z

BMC Genomics

RESEARCH ARTICLE Open Access

Analysis of whole genome sequencing for the *Escherichia coli* O157:H7 typing phages

Lauren A Cowley[1*], Stephen J Beckett[2], Margo Chase-Topping[3], Neil Perry[1], Tim J Dallman[1], David L Gally[3] and Claire Jenkins[1]

Abstract

Background: Shiga toxin producing *Escherichia coli* O157 can cause severe bloody diarrhea and haemolytic uraemic syndrome. Phage typing of *E. coli* O157 facilitates public health surveillance and outbreak investigations, certain phage types are more likely to occupy specific niches and are associated with specific age groups and disease severity. The aim of this study was to analyse the genome sequences of 16 (fourteen T4 and two T7) *E. coli* O157 typing phages and to determine the genes responsible for the subtle differences in phage type profiles.

Results: The typing phages were sequenced using paired-end Illumina sequencing at The Genome Analysis Centre and the Animal Health and Veterinary Laboratories Agency and bioinformatics programs including Velvet, Brig and Easyfig were used to analyse them. A two-way Euclidian cluster analysis highlighted the associations between groups of phage types and typing phages. The analysis showed that the T7 typing phages (9 and 10) differed by only three genes and that the T4 typing phages formed three distinct groups of similar genomic sequences: Group 1 (1, 8, 11, 12 and 15, 16), Group 2 (3, 6, 7 and 13) and Group 3 (2, 4, 5 and 14). The *E. coli* O157 phage typing scheme exhibited a significantly modular network linked to the genetic similarity of each group showing that these groups are specialised to infect a subset of phage types.

Conclusion: Sequencing the typing phage has enabled us to identify the variable genes within each group and to determine how this corresponds to changes in phage type.

Background

Escherichia coli O157:H7 is the most prevalent Shiga toxin producing *E. coli* (STEC) serotype in the UK and has the most severe impact on human health [1]. STEC O157 symptoms can range from mild gastroenteritis to severe bloody diarrhoea and in more extreme cases haemolytic uraemic syndrome (HUS) [2]. The very young, elderly and immune-compromised are particularly at risk of HUS. A recent Public Health England (PHE) study found incidence to be as high as 1.78 per 100,000 person-years with up to 33% of cases being hospitalised (Gastrointestinal Bacterial Reference Unit (GBRU) in house data). The GBRU at PHE receives approximately 1000 STEC O157 samples per year. Recent outbreaks in the UK have been foodborne or linked to petting farms [3-5]. For purposes of public health surveillance and outbreak investigations, STEC strains are differentiated by phage typing and multilocus variable number tandem repeat analysis [6].

Bacteriophages are viruses that infect bacteria and cause bacterial lysis and cell death, but can also promote horizontal gene transfer between bacteria, play an important role in dynamic bacterial genome evolution and can regulate the abundance and diversity of bacterial communities through co-evolution [7]. There are a range of phages that infect *Escherichia coli* that progress either to a lytic or lysogenic phase after infection. A lytic phase will cause cell lysis whereas in lysogenic phase the phage becomes integrated into the host genome and becomes a prophage. Prophages are important as they often encode additional factors not directly linked to phage production that may provide an evolutionary advantage to the bacterial host enabling survival of the embedded prophage. These include factors that promote colonisation of animal hosts as well as their regulators [8,9]. Bacteriophage specificity is, in part, dependent on

* Correspondence: lauren.cowley@phe.gov.uk
[1]Gastrointestinal Bacteria Reference Unit, Public Health England, 61 Colindale Ave, London NW9 5HT, UK
Full list of author information is available at the end of the article

BioMed Central

© 2015 Cowley et al.; licensee BioMed Central. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

the ability of tail fiber proteins to bind to specific receptors on the bacterial host [10].

Phage-typing of STEC O157 is a scheme based on the use of 16 bacteriophages that produce a phage infection profile for a strain based on the level of lysis achieved by each phage [11] and has been used to categorize outbreaks and sporadic cases. Today 80% of all STEC O157 strains typed are PT 8, 21/28, 2, 4 or 32 in the UK (GBRU in house data). Certain PTs are more likely to be associated with human infection and so far there is little understanding of the basis for this. While ongoing work is focused on sequencing and analysis of the bacterial strains, we propose that further insight into relevant strain differences can be gained by also understanding the typing phages themselves and the basis of their infection selectivity. A longer term aim of the work is to understand the factors that mediate resistance and susceptibility in the phage-bacterium relationship.

Little is known about the molecular basis for the interaction between phages and different strains of different phage types, however we can interrogate the phage infection profile of who-infects-whom as a bipartite (two-mode) network. Two common methods for analysing community structure in bipartite data are nestedness and modularity. Nestedness is a way of measuring the ranges of both host resistance and phage infectivity across a specialist to generalist gradient. Specialists are assumed to have strategies that are subsets of those which are more generalised. Modularity is the degree to which a network can be split into distinct modular groupings of phage and bacteria such that there are many infections within rather than between groups [12].

The 16 phages in the STEC phage-typing scheme are made up of 14 T4 phages and 2 T7 phages. An example of a T7 phage has been sequenced previously and T7 are known to consist of a single 'chromosome' carrying about 30 genes [13]. The 5' end genes of the chromosome are expressed at an early stage of infection and their products are involved in the induction of host RNA polymerase for transcription and control the expression of other phage genes in a positive feedback mechanism. Genes that are expressed later are involved in the metabolism of phage DNA and code for capsid proteins or are involved in the assembly of infective progeny particles [13]. T4 phages have much larger genomes with 300 putative genes, only 62 of these have been found to be 'essential' under laboratory conditions [14]. The order of expression works in a similar way to T7 phage.

The STEC O157 typing phages 5, 7 and 10 from the typing scheme have previously been sequenced [15-17]. Our sequencing results are consistent with previously published sequences. We build on this data by placing the previously sequenced phages into similarity groups within the typing phages. The aim of this study was to analyse the genome sequences of 16 (fourteen T4 and two T7) STEC O157 typing phages (TPs) and to identify genes that may account for differences in infectivity between related phages.

Methods

Phage propagation and DNA extraction

The typing phages were obtained as a gift from the National Microbiology Laboratory, Winnipeg, MN, Canada to GBRU in the late 1980s. To propagate the phage, 0.1 ml of the propagating strain (Additional file 1: Table S1, Figure 1) was inoculated into 2 × 20 ml of single strength Difco nutrient broth and 0.1 ml of test phage was added to one and the other kept as a control. The bottles were incubated and turbidity was monitored. When lysis was judged to be at its maximum compared to the control, a small amount of the phage solution was centrifuged at 2,200 g for 20 min. The supernatant was removed and spotted onto a flooded plate of propagating strain as a test; the plate was dried and incubated at 37°C overnight. The plates were examined for lysis and if positive the phage lysate was sterilized by filtration and stored at 4°C.

All phages were filtered before extraction took place. Eleven (phages 1, 3, 4, 5, 6, 7, 8, 9, 12, 13 and 14) of the 16 phages were extracted using the QIAamp UltraSens Virus kit (Qiagen, UK) following the manufacturer's instructions. This method failed to produce a high enough concentration of DNA for the remaining phages (2, 10, 11, 15 and 16) and these were extracted using a Zinc Chloride protocol [18]. Briefly, 20 μl of a 2 M Zinc chloride solution was added to 1 ml of sample and incubated for 5 min at 37°C. The sample was then centrifuged at 10000 rpm and the supernatant was removed. The pellet was resuspended in 500 μl of TES buffer (0.1 M Tris–HCl, pH8; 0.1 M EDTA and 0.3% SDS) and then incubated at 60°C for 15 min. Subsequently, 60 μl of a 3 M potassium acetate solution was added and the sample left on ice for 10 to 15 min. Following the formation of a white, dense precipitation the sample was centrifuged for 1 min at 12000 rpm and the supernatant removed to a new tube. To this an equal volume of isopropanol was added, the solution vortexed and left on ice for 5 min. The solution was centrifuged and evaporated simultaneously using a Speedy-Vac machine and the pellet washed with 70% ethanol before being resuspended in 20–100 μl TE (10 mM Tris–HCl, pH8; ImM EDTA). Samples were pooled by five extractions to give a higher yield of DNA. This method also failed to produce high enough concentration of DNA for sequencing TP 2 and 16 and we were ultimately unable to obtain sequencing data for these two TPs.

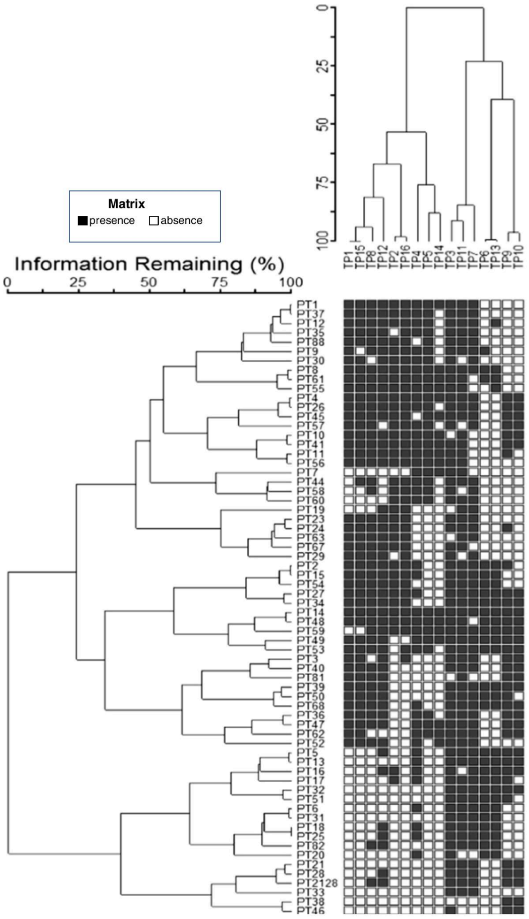


Figure 1 Two-way cluster analysis dendrogram of 66 phage types and 16 typing phages. The matrix of shaded squares represents the phage type × typing phage matrix, while the dendrograms show the clustering. The dendrograms are scaled by Wishart 's (1969) objective function, expressed as the percentage of information remaining at each level of grouping (McCune and Grace, 2002). Each square represents the presence (black) and absence (white) of a reaction with a given typing phage. The three phage type clusters and the 4 typing phage clusters are indicated at the node with numbers.

Sequencing

The first set of phages (1, 3, 4, 5, 6, 7, 8, 9, 12, 13 and 14) was sequenced at The Genome Analysis Centre (TGAC) on an Illumina MiSeq. Illumina TruSeq DNA library construction was performed and sequencing of the libraries was pooled on one run using 150 bp paired-end reads, this generated greater than 1 Gbp of data for the run. Data was then quality controlled, basecalling was performed and it was formatted. The second set of phages (10, 11 and 15) was sequenced at the Animal Health and Veterinary Laboratories Agency on an Illumina GAII. The library construction was performed using a Nextera DNA sample preparation kit (Illumina) and then sequenced in the same manner as the other set.

Bioinformatic sequencing analysis

Reads for all phages apart from TP 15 were *de novo* assembled into whole genomes using Velvet optimizer with a range of k-mer values from 90–120 [19] and annotated using Prokka 1.5.2 and output as GenBank files [20]. The genomes were visualised in the multiple genome alignment tool Mauve with a progressive alignment to visualise similarities and differences between them based on sequence content. The reads assembled into between 1 and 7 contigs for each phage.

TP15 could not be assembled correctly because the propagation process had induced other temperate phages in the genome of the propagating strain and the DNA had been co-extracted. Subsampling to x150 coverage and the genome assembler SPAdes with a better low frequency k-mer elimination step [21] was used to overcome this issue and resolve 15 true typing phage 15 contigs from the assemblies. The sequencing data has been made publicly available in the Short Read Archive under study alias PRJNA252693 and Genbank accession numbers for each phage can be found in the availability of supporting data section.

Euclidian tree

Data from PHE on the protocol used to identify phage types (Additional file 1: Table S3, Additional file 1: Table S2) was converted into binary (presence/absence) format. In the original scheme there were 66 established phage types (PT) and 16 typing phages (TP). This set of data was analysed using a two-way cluster hierarchical agglomerative analysis in PC-ORD software version 6.08 (MJM software Design, Gleneden Beach, OR). The clustering was performed with Euclidian distance matrix and Ward linkage method.

The optimal number of groups of plots was first evaluated with multiresponse permutation procedure, seeking the solution with fewest number of groups but the greatest gain in *A*-statistics [22].

Modularity and nestedness

Modularity of the network was calculated using the LPAb + algorithm [23] which uses label propagation coupled with greedy multistep agglomeration to identify the communities (made of members of both types of nodes (bacteria and phage)) that maximise modularity in bipartite networks. As LPAb + is stochastic we choose the best modularity score, Q_B, returned from 1,000 trials each time we use the algorithm. Code for performing the modularity analysis is supplied [24].

Nestedness statistics were calculated using FALCON [25]. The nestedness measures used were NODF [26], NTC [27,28] and BR, the discrepancy score of Brualdi and Sanderson, 1999 [29]. NODF and NTC scores take values in the range [0,100], whilst BR is the absolute number of differences between the input and a maximally packed matrix. NODF has been recalculated here as NODF = 100-NODF, so that lower measure scores show greater nestedness with 0 representing perfect nestedness for each of the measures.

We tested for significance of both modularity and the nestedness found in our phage-bacteria infection network using two null models based on properties of our network. Null model one is a Bernoulli random null model where connections between phage j and bacteria i are made with probability $p_{ij} = F/M$, where F is the total number of edges in our network (number of infecting interactions) and M is the maximum number of potential interactions (number of TP's × number of PT's). Null model two is based on the information in the rows and columns in the network [30]; where a connection between phage *j* and bacteria *i* is made with probability $p_{ij} = 0.5\ (d_j/r + k_i/c)$ where d_j is the number of infections caused by phage *j*, *r* is the number of PTs, k_i is the number of phage that can infect bacteria *i* and *c* is the number of TPs. We tested 1,000 null matrices against our network for each null model in the modularity analysis, whilst we used the adaptive ensemble of FALCON for nestedness analysis and report the ensemble size used (N), p-values (probability of finding a more modular/nested network from the null model) and z-scores (effect size; the number of standard deviations our network was away from the mean average found in each null model).

BRIG plot

BRIG (Blast Ring Image Generator), a genome comparison tool [31], was used to compare similarities between the 12 T4 like typing phages by inputting all of the GenBank files for the assembled genomes and plotting blast hits against a MultiFASTA file of all of the phages. The image was displayed as a series of concentric rings with the central ring being the MultiFASTA reference; each outer ring displays hits (i.e. genomic regions that show a high percentage similarity to the central reference genome) for each phage. BRIG was also used to show the comparison of phages 9 and 10 (the two T7 like typing phages) against phage 9 as a reference.

SeqFindR and Easyfig plots

SeqFindR, a bioinformatics tool developed by the Beatson Laboratory at the University of Queensland, was used to identify gene presence and absence in the phage genomes. Easyfig [32] was used to visualise the coding regions and colour the accessory genes in red for each phage group.

Tail fiber analysis

Tail fiber encoding genes were extracted from the GenBank files of the typing phages and the protein sequences aligned using MEGA 5.2. The alignment told us how many changes in protein sequence there were within the groups.

Results

In the phage typing scheme there are 14 T4-like bacteriophages (TP1-8 and TP11-16) and two T7-like bacteriophages (TP9 and TP10). The reactivity of each of the typing phages with respect to the STEC O157 phage typing scheme was analysed. The two-way Euclidian cluster analysis combined the independent clustering of 66 STEC O157 bacterial phage types and the 16 typing phages into a single diagram and highlighted the associations between groups of phage types and typing phages (Figure 1). The analysis showed that the STEC O157 phage typing scheme formed a weak ($Q_b = 0.1575$ (Table 1)) but significantly modular network where the TP groups were each specialised to infect a subset of PTs (Figure 2). There also exists a large number of between module interactions. Furthermore, the majority of PTs of STEC O157 react with at least one member of each group of typing phages. These groups can be regarded as universally infective against STEC O157. Using statistical tests we also found that the nestedness of our interaction network was statistically significantly different from that found under randomly formed networks (Table 1). This indicates a correlation between phage infectivity range and the resistance range of the host. These phages have been designed and chosen to infect STEC O157 and create a typing scheme with the simplest and minimum selection of phages so it makes sense that the system is nested.

Table 1 Summary statistics for nestedness and modularity analysis

		Modularity	Nestedness		
Measure		Q_B	NODF	NTC	BR
Measure score	x	0.1575	27.9199	30.2532	130
Null model 1	N	1000	1300	1300	1300
	p-value	<1/N	<1/N	<1/N	<1/N
	z-score	4.8602	-7.5382	-11.9831	-11.7632
Null model 2	N	1000	1000	1000	1000
	p-value	<1/N	<1/N	<1/N	<1/N
	z-score	5.7693	-4.6740	-6.7842	-7.1554

Barber's modularity (Q_b) and three nestedness measures (NODF, NTC and BR) were calculated. Two null models were used to generate ensembles of networks (of size N) to evaluate the strength of the modularity and nestedness observed in the classified *Escherichia coli* O157:H7 phage-bacteria infection network. This is done by reporting the significance (as a *p*-value) and effect size (as a z-score) of the phage-bacteria infection network relative to the networks found in each null model ensemble. Note that, due to differences in how these measures are calculated, for modularity a positive z-score indicates that modularity is greater in the observed network than the mean average of the ensemble; whilst in the nestedness analysis a negative z-score indicates the observed network is more nested than the mean nestedness found within the null ensemble. The classified *Escherichia coli* O157:H7 phage-bacteria infection network was found to be both more nested and more modular than any of the networks generated by the tested null models.

Fourteen of the 16 phages in the typing scheme were sequenced and successfully assembled. Despite several attempts, sequencing of typing phages 2 and 16 failed due to insufficient quantities of DNA extracted from the phage propagation preparations.

The BRIG plot showed that the 12 sequenced T4-like bacteriophages formed three distinct groups of similar genomic sequences (Figure 3). Group 1 included typing phages 1, 8, 11, 12 and 15; Group 2 comprised typing phages 3, 6, 7 and 13 and typing phages 4, 5 and 14 were in Group 3. Although the sequencing for TP2 and TP16 failed, the modularity analysis indicates that TP16 belonged to Group 1 and TP2 belonged to Group 3 (Figure 2). The TPs varied significantly in size between the three groups: the members of Group 1 were 93,000–95,000 bp, Group 2 members were 165,000–175,000 bp and those in Group 3 were 135,000–140,000 bp.

The Group 1 phages (TP1, 8, 11, 12 and 15) were approximately 90,000 bp in length. These five phages were highly similar in genetic sequence content. The location, annotation and presence of accessory genes within Group 1 are shown in Figure 4, Additional file 1: Table S3. Figure 4 shows that there were 6 genes found in TP1 but absent in TP8, 11, 12 and 15 (five were annotated as hypothetical proteins and one tRNA). There were also five genes present in TP8, 11, 12 and 15 but not in TP1 (three were annotated as hypothetical proteins, one as AP2 domain protein and one was a tRNA gene) (Figure 4, Additional file 1: Table S3). TP8 was missing a region annotated as a putative prophage that was present in the other members of the group. With the exception of TP11, the Group 1 TPs are most closely related to each other by the two-way Euclidian cluster analysis demonstrating the link between gene content and phage typing profile.

The typing phages in Group 2 (TP 3, 6, 7, and 13) were between 160–170,000 bp in length. The genomes were almost twice the size of the phages in Group 1 and exhibited less similarity. The accessory genes found in Group 2 were mostly annotated as encoding hypothetical proteins (Figure 5, Additional file 1: Table S4). The two-way Euclidian cluster analysis highlighted a close relationship between TP6 and TP13 and this corresponded with the level of sequence similarity of these two typing phages illustrated in Figure 5.

Typing phages 4, 5 and 14 were designated Group 3 and were 130–140,000 bp in length. Figure 6 shows the location, annotation and presence of accessory genes within Group 3. Figure 6 demonstrates that there were 29 gene differences within the group and the majorities

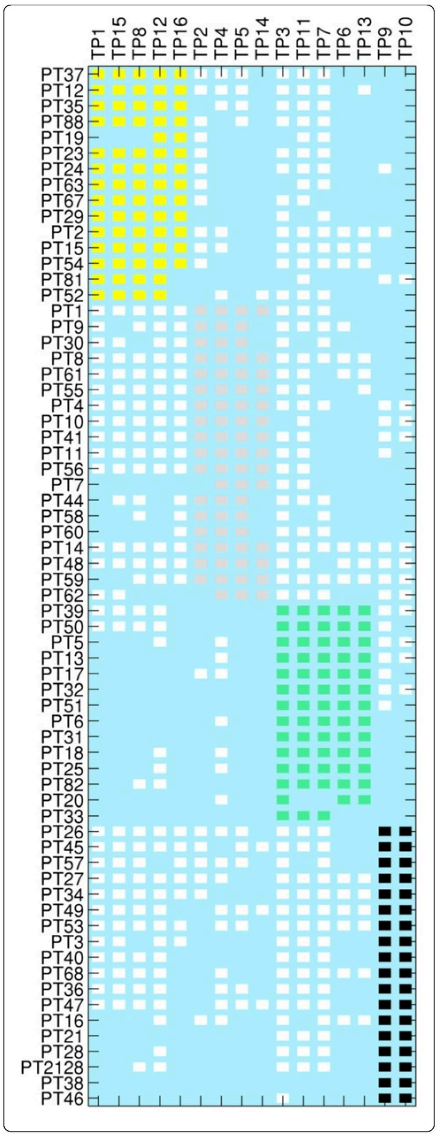


Figure 2 A visual representation of the modularity seen within the system with modules coloured. Phage type (PT) is represented on the y axis and Typing phage (TP) is represented on the x axis and the matrix showing presence of a reaction with that phage as a white or coloured block. The 4 observed modules are coloured as yellow, pink, green and black.

(19) were annotated as hypothetical proteins. In addition, three genes encoded putative endonucleases and there were three genes designated tRNAs (Figure 6, Additional file 1: Table S5). The typing phages in Group 3 were most closely related to each other by the two-way Euclidian cluster analysis (Figure 1).

Phages 9 and 10, the two Podoviridae or T7 like phages that are found in the typing scheme, were successfully sequenced, assembled and annotated and revealed 40–45000 bp genomes consistent with the published sequences of T7 bacteriophages (Figure 7). Phages 9 and 10 differed by only three genes (annotated as encoded hypothetical proteins) that were found in Phage 9 but not in phage 10. The two-way Euclidian cluster analysis confirmed the close relationship between TP9 and TP 10 in terms of phage type profile. It also showed that there were six STEC O157 phage types (PT 2, 11, 17, 24, 50, and 51) that react with TP9 but not TP10 and none of the phage types react with TP10 but not TP9 (Figure 1). These three hypothetical proteins could be the key to the differences in the reactivity profiles of TP9 and 10.

Tail-fiber encoding genes were analysed within each group and it was found that there were changes in the amino acid sequence for certain members of each group. Within the group 1 typing phages, phages 1 and 15 had 3 changes in amino acid sequence in their tail fibers, 2 of which were shared and 1 each unique to each phage. Within the group 2 typing phages, phage 7 has 47 changes in its amino acid sequence and 3 amino acid deletions. Within the group 3 typing phages, the same single position in all 3 members of the group has a different amino acid present and additionally there was another single position in typing phage 14 that had a different amino acid. The T7-like phages had identical tail fiber genes. There was no genetic similarity between tail fiber genes found in different groups.

Discussion

Phage-host interactions are key to understanding the virulence and success of *E. coli* O157 but little is known about the typing phages used in the O157 typing scheme. Sequencing these phages has enabled us to group the T4-like Myoviridae and the two Podoviridae or T7-like phages members of the typing phage scheme into four groups based on their sequence similarity. The two-way Euclidian cluster analysis demonstrated that

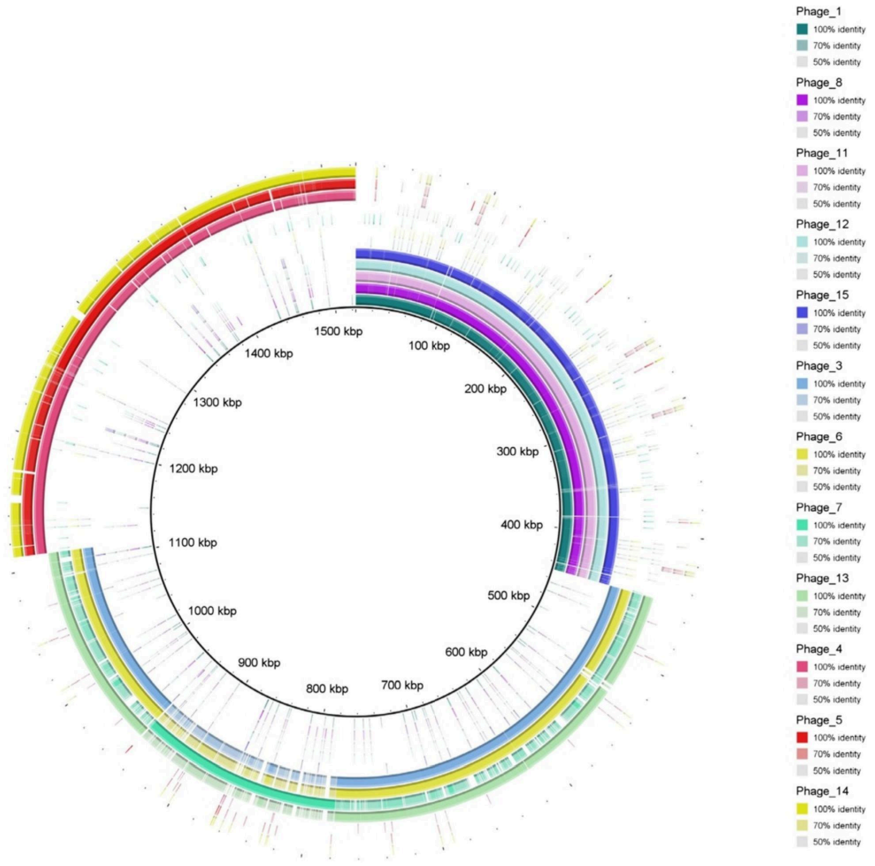


Figure 3 A genomic representative diagram drawn with BRIG of T4-like phage similarities, the coloured regions indicate high pairwise genomic sequence similarity according to blastn. Legend indicates which colours correspond with which phages and the shade of that colour indicates what level of similarity is observed. Central ring is multifasta of all T4-like phage genomes and each consecutive ring represents the similarity with a single phage. The multifasta and rings are in the same phage order.

similar phage groups react with STEC O157 in a similar way with closely related reaction profiles. The sequencing data also highlighted that a small number of gene differences may be responsible for the subtle differences in reaction profiles within the groups.

The large proportion of genes annotated as encoding hypothetical proteins hindered our investigations into the mechanisms of host-phage interactions. Attempts were made to annotate these genes further using protein BLAST and HMMER but only uncharacterised proteins were hit. However, the determination of which genes vary within each group will enable us to focus on the genes that may play a key role in the mechanisms of interactions between specific typing phages and strains belonging to specific phage types. For example, in Group 1, there were five genes that were found only in TP8, 11 and 12 and three PTs (PT21/28, 59 and 82) that only react with these TPs. The proteins encoded by these five

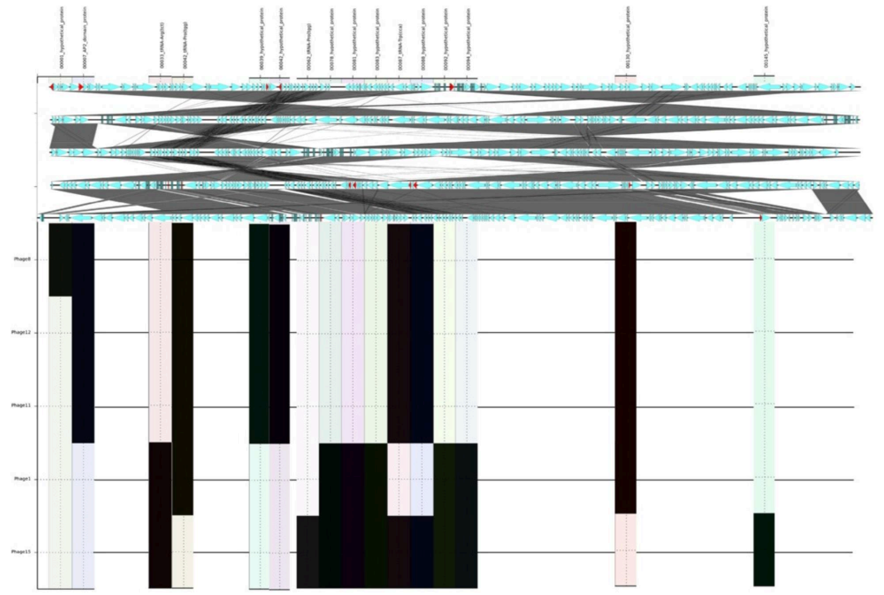

00003_hypothetical_protein
00007_AP2_domain_protein
00033_tRNA-Arg(tct)
00042_tRNA-Pro(tgg)
00039_hypothetical_protein
00042_hypothetical_protein
00042_tRNA-Pro(tgg)
00078_hypothetical_protein
00081_hypothetical_protein
00083_hypothetical_protein
00087_tRNA-Trp(cca)
00088_hypothetical_protein
00092_hypothetical_protein
00094_hypothetical_protein
00130_hypothetical_protein
00145_hypothetical_protein
Phage8
Phage12
Phage11
Phage1
Phage15

Figure 4 SeqFindR and Easyfig image combined representing the accessory gene content of group 1. Genomes of each phage in group 1 are represented by the Easyfig image showing linear visualisation of the genome and coding regions represented by arrows, accessory genes are coloured red. The order of phage genomes in the linear visualisation and the accessory content blocking is 8, 12, 11, 1 and 15 and was chosen based on similarity clustering in SeqFindR. Hits for the accessory genes in each genome are represented in labelled columns in the SeqFindR image underneath each accessory gene.

genes may play a key role in the host-phage mechanisms between TP8, 11 and 12 and strains of STEC O157 belonging to PT21/28, 59 and 82. PT21/28 is the most common PT in the UK and is significantly associated with HUS [33]. Further details of unique host-phage interactions are described in Additional file 1: Table S6 and the genes referred to within can be found in Figures 4, 5 and 6.

Analysis of tail fibers genes showed that typing phages 1, 15, 7 and each individual member of Group 3 had different protein sequences encoded to the other members of their group. The changes that were found could partially account for infectivity differences [34]. These could explain a few of the differences in host specificity seen within those groups, although this will not apply to the T7-like typing phages that have had identical predicted tail fiber proteins.

Certain typing phages had almost identical genomes but different host susceptibility profiles, for example, TP11 belonged to the Group 1 phages but had a similar host susceptibility profile to the Group 2 phages. Each phage in the typing scheme has its own propagating strain (see Additional file 1: Figure S1, Table 1) so it is also possible that host-induced modification occurs [35]. For example, the propagating strains for the closely related typing phages TP9 and TP10 are STEC O157 PT2 and PT32, respectively. Modifications may be a result of methylation or other phenotypic changes that are not evident in the genome but may affect the host range of the virus.

Phenotypic differences in susceptibility patterns in genetically similar phages could be explained by the transcription order of genetic loci in the phage genome. It has been suggested that gene synteny constrains adaptation and is important for fitness and, therefore, infectivity of bacteriophages [36]. The order of transcription may be important in overcoming the host response to infection. The phages that transcribe their genetic loci in a different order may be killed and degraded by the host response, for example, TP 8, 11 and 12 are almost identical but have a different gene order and this may be key to their different infection profiles.

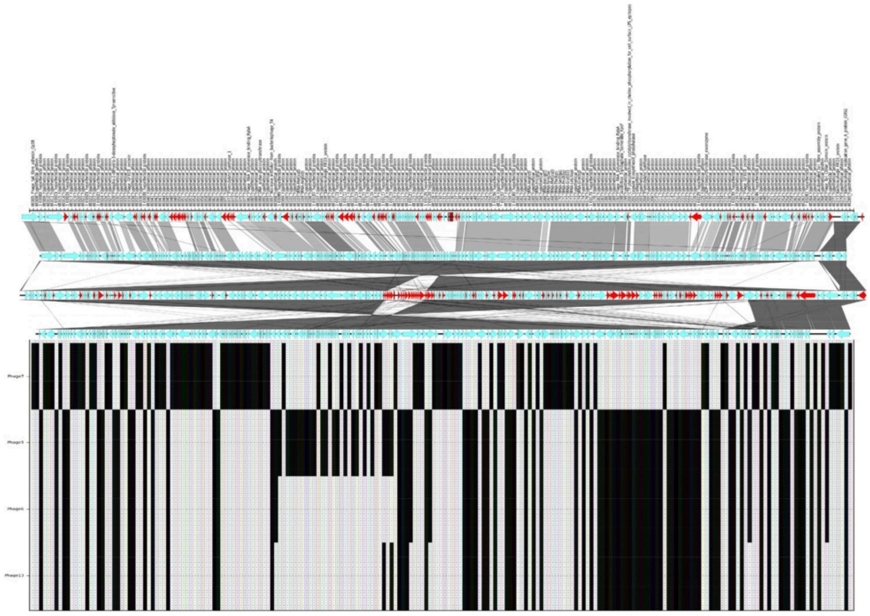


Figure 5 SeqFindR and Easyfig image combined representing the accessory gene content of group 2. Genomes of each phage in group 2 are represented by the Easyfig image showing linear visualisation of the genome and coding regions represented by arrows, accessory genes are coloured red. The order of phage genomes in the linear visualisation and the accessory content blocking is 7, 3, 6 and 13 and was chosen based on similarity clustering in SeqFindR. Hits for the accessory genes in each genome are represented in labelled columns in the SeqFindR image underneath each accessory gene.

Our analysis showed that the significantly modular network exhibited by the STEC O157 phage typing scheme was linked to the genetic similarity groups mentioned above showing that these groups are specialised to infect a subset of PTs. However, the typing scheme as a whole is also significantly nested; more generalised phages minimise the number of phages needed in the scheme. Both of these network structures have also been found in other phage-bacteria infection networks [37,38]. The most common PTs in the UK: 2, 8, 21/28 and 32 are all found in different modules, meaning there is an abundant PT in each module. When looking at these PTs with nestedness, PT 8 and 2 both have a phage susceptibility range of 14 and 13 respectively so are quite generalised but PT 21/28 and 32 both have a host range of 7, and lie more towards the specialised end of the spectrum. It is interesting that the more abundant strains seem to appear at two levels of host range – perhaps suggestive of a trade-off between host range and phage productivity. It would be interesting to see, in conditions where the phages are allowed to evolve with their hosts, if a more modular network arises with further specialisation of the phages to maintain a kill-the-winner dynamic and less broad range infectivity [39]. This is an artificial system that we are observing and it is likely that we would see a different network arising in nature's ecological systems.

Phage-typing has been used for epidemiological and surveillance studies by a number of groups [40,41] for different organisms. Phage-type association with increased strain virulence is of high interest to public health workers dealing with STEC O157, the replacement of phage-typing with whole genome sequencing should still incorporate our knowledge of phage type and associated virulence. For this reason it is valuable to find the molecular markers associated with high

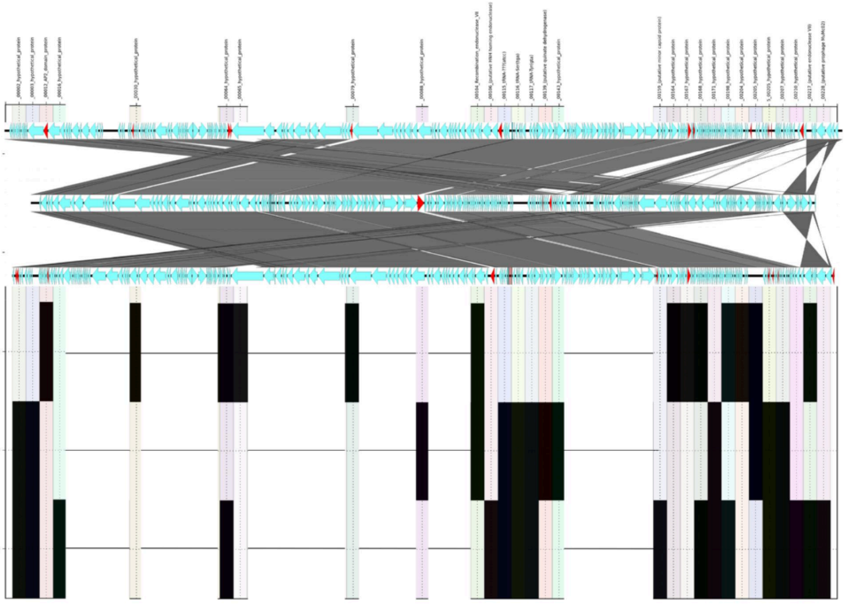

_00002_hypothetical_protein
_00003_hypothetical_protein
_00012_AP2_domain_protein
_00016_hypothetical_protein
_00030_hypothetical_protein
_00064_hypothetical_protein
_00065_hypothetical_protein
_00079_hypothetical_protein
_00088_hypothetical_protein
_00104_Recombination_endonuclease_VII
_00106_(putative HNH homing endonuclease)
_00116_tRNA-Ser(tga)
_00117_tRNA-Tyr(gta)
_00139_(putative quinate dehydrogenase)
_00143_hypothetical_protein
_00159_(putative minor capsid protein)
_00164_hypothetical_protein
_00167_hypothetical_protein
_00168_hypothetical_protein
_00171_hypothetical_protein
_00198_hypothetical_protein
_00204_hypothetical_protein
_00205_hypothetical_protein
_00207_hypothetical_protein
_00210_hypothetical_protein
_00217_(putative endonuclease VII)
Phage4
Phage14
Phage5

Figure 6 SeqFindR and Easyfig image combined representing the accessory gene content of group 3. Genomes of each phage in group 3 are represented by the Easyfig image showing linear visualisation of the genome and coding regions represented by arrows, accessory genes are coloured red. The order of phage genomes in the linear visualisation and the accessory content blocking is 4, 14 and 5 and was chosen based on similarity clustering in SeqFindR . Hits for the accessory genes in each genome are represented in labelled columns in the SeqFindR image underneath each accessory gene.

frequency and highly pathogenic phage types; elucidating the determinants underpinning differences in phage typing should contribute to this.

Phage-mediated therapies will continue to be an area of interest as we struggle with resistance to conventional antibiotics. It makes sense that moving forward there will be considerable interest in being able to predict bacterial susceptibility to 'treatment' phages based on sequence information alone. Furthermore, the next step would be modification of specific phages to improve their targeting/activity. This will rely on understanding of the phage genes that govern the specificity of infection in different backgrounds. The place to start is with certain key bacterial pathogens and a bank of phages.

Conclusions

In this study, the STEC O157 typing phages we clustered into four distinct groups of similar genomic sequences, that broadly correlated with phage typing profile groups determined by two-way Euclidian clustering. Genetic variation within the TP groups may explain the subtle differences between the phage typing profiles exhibited by the *E. coli* O157 typing phages. This analysis was hindered by the lack of detailed annotation of protein encoding genes in T4 and T7-like phages. The impact of the order of transcription of the blocks of genetic loci and the role of host-induced modification further confound the analysis. However, sequencing the typing phage has enabled us to identify the variable genes within each group and to determine how these correspond to changes in phage type. Future studies will focus on the genes that appear to alter host-phage interactions and we aim to identify bacterial genes that influence typing phage resistance and susceptibility using random mutagenesis approaches. In order to understand the best combination of strains and individual phages to work with, the network of interactions needs to be analysed. This information can also provide insight on how phage

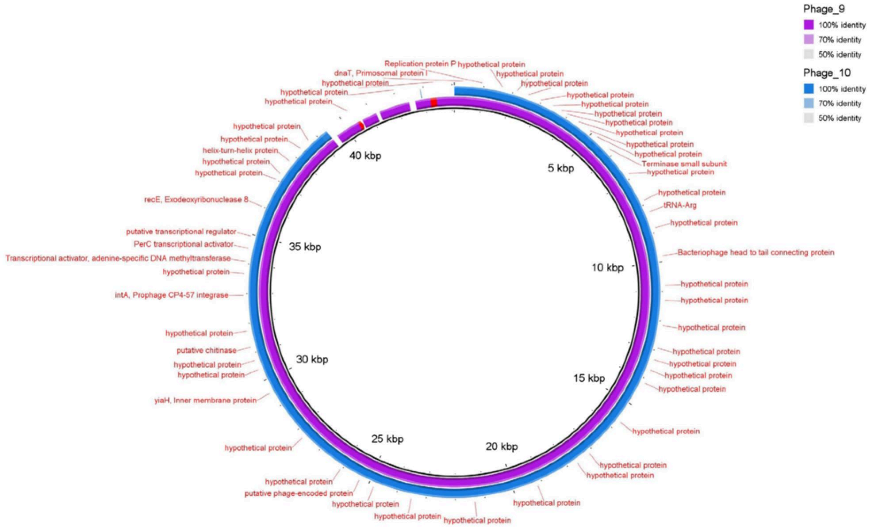


Figure 7 A genomic representative diagram drawn with BRIG, the coloured regions indicate high pairwise genomic sequence similarity according to blastn. The legend indicates which colours correspond with which phages. The central ring is a genbank file of Phage 9 as a reference and annotations of genes in red. The first ring is representative of Phage 9 and the second ring is representative of Phage 10 and the shade indicates the level of genomic similarity observed.

typing can potentially be simplified in the future. A better understanding of the genetic differences between bacterial phage types, and the possible differences in virulence factors, could help elucidate why different phage types occupy specific niches and are associated with different patient age groups and disease severity.

Availability of supporting data

The raw sequencing reads have been deposited in the short read archive under project alias PRJNA252693. The assembled sequences and annotations of the typing phages have been deposited in Genbank under the following accessions;

Phage 1: KP869100
Phage 3: KP869101
Phage 4: KP869102
Phage 5: KP869103
Phage 6: KP869104
Phage 7: KP869105
Phage 8: KP869106
Phage 9: KP869107
Phage 10: KP869108
Phage 11: KP869109
Phage 12: KP869110
Phage 13: KP869111
Phage 14: KP869112
Phage 15: KP869113

All other supporting data is included as additional files.

Additional file

Additional file 1: Table S1. Propagating strain table. Table showing propagating strain and corresponding typing phage number that the strain propagates. **Table S2.** *E. coli* O157 phage typing scheme. Table showing reactions of the *E. coli* O157 type strains with the typing phages. **Table S3.** Table of the accessory variation of the Group 1 typing phages. Table detailing the accessory variation of Group 1 as depicted in figure 4. **Table S4.** Table of the accessory variation of the Group 2 typing phages. Table detailing the accessory variation of Group 2 as depicted in figure 5. **Table S5.** Table of the accessory variation of the Group 3 typing phages. Table detailing the accessory variation of Group 3 as depicted in figure 6. **Table S6.** Table of unique reactions. Table representing unique reaction that only occur within a subset of groups 1, 2 and 3 with specific PTs and number of genes found only in that subset. **Figure S1.** Phylogenetic tree of propagating strains. Phylogenetic tree of propagating strains for each typing phage and sakai as a reference. **Figure S2.** Visual representation of nestedness. A visual representation of the degree of nestedness found within the classified *E. coli* O157 phage-bacteria infection network. **Figure S3.** Electron microscopy image of typing phage 7 A representation of the the T4-like long-tailed phage morphology within the typing phages. **Figure S4.** Electron microscopy image of typing phage 9 A representation of the T7-like short-tailed phage morphology within the typing phages.

Abbreviations

STEC: Shiga Toxin producing *Escherichia coli*; HUS: Haemolytic uraemic syndrome; PT: Phage-type; TP: Typing phage; TGAC: The genome analysis centre; AHVLA: Animal health and veterinary laboratories agency; PHE: Public health England; GBRU: Gastrointestinal bacterial reference unit; BRIG: Blast ring image generator.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

LC carried out the genetic analysis of the typing phages and drafted the manuscript. SB carried out the modularity/nestedness analysis. MCT performed the statistical analysis necessary to draw the two-way Eculidian tree. NP propagated the typing phages. TJD helped to draft the manuscript. DLG helped to draft the manuscript. CJ conceived the wider study and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Matthew Hannah for performing the electron microscopy needed to take the pictures of the bacteriophages, Richard Ellis for performing sequencing of the bacteriophages at AHVLA, Vivienne do Nascimento for propogating the bacteriophage and Phillip Ashton for bioinformatics support. LAC acknowledges PhD funding from Public Health England and the National Institute for Health Research scientific research development fund (108601). DLG is supported by BBSRC institute strategic grant funding from the BBSRC.

Author details

[1]Gastrointestinal Bacteria Reference Unit, Public Health England, 61 Colindale Ave, London NW9 5HT, UK. [2]Biosciences, College of Life and Environmental Sciences, University of Exeter, Laver Building, North Park Road, Exeter EX4 4QE, UK. [3]Division of Immunity and Infection, The Roslin Institute, R(D)VS, University of Edinburgh, Edinburgh EH25 9RG, UK.

Received: 16 July 2014 Accepted: 17 March 2015
Published online: 08 April 2015

References

1. Adak GK, Long SM, O'Brien SJ. Trends in indigenous foodborne disease and deaths, England and Wales: 1992 to 2000. Gut. 2002;51:832–41.
2. Griffin PM, Ostroff SM, Tauxe RV, Greene KD, Wells JG, Lewis JH, et al. Illnesses associated with Escherichia coli O157:H7 infections. A broad clinical spectrum. AnnInternMed. 1988;109:705–12.
3. Underwood AP, Dallman T, Thomson NR, Williams M, Harker K, Perry N, et al. Public health value of next-generation DNA sequencing of enterohemorrhagic Escherichia coli isolates from an outbreak. JClinMicrobiol. 2013;51:232–7. doi:JCM.01696-12; 10.1128/JCM.01696-12.
4. Perry N, Cheasty T, Dallman T, Launders N, Willshaw G. Application of multilocus variable number tandem repeat analysis to monitor Verocytotoxin-producing Escherichia coli O157 phage type 8 in England and Wales: emergence of a profile associated with a national outbreak. JApplMicrobiol. 2013;115:1052–8. doi:115:1052–1058; 10.1111/jam.12303.
5. Ihekweazu C, Carroll K, Adak B, Smith G, Pritchard GC, Gillespie IA, et al. Large outbreak of verocytotoxin-producing Escherichia coli O157 infection in visitors to a petting farm in South East England, 2009. EpidemiolInfect. 2012;140:1400–13. doi:S0950268811002111;10.1017/S0950268811002111.
6. Byrne L, Elson R, Dallman TJ, Perry N, Ashton P, Wain J, et al. 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. PLoSOne. 2014;9:e85901. doi:10.1371/journal.pone.0085901; PONE-D-13-31410.
7. Brockhurst MA, Koskella B. Experimental coevolution of species interactions. Trends EcolEvol. 2013;28:367–75. doi:S0169-5347 (13) 00061-X;10.1016/j.tree.2013.02.009.
8. Hoey DE, Currie C, Else RW, Nutikka A, Lingwood CA, Gally DL, et al. Expression of receptors for verotoxin 1 from Escherichia coli O157 on bovine intestinal epithelium. J Med Microbiol. 2002;51:143–9.
9. Ohnishi M, Kurokawa K, Hayashi T. Diversification of Escherichia coli genomes: are bacteriophages the major contributors? Trends Microbiol. 2001;9:481–5. doi:S0966-842X (01) 02173-4.
10. Scholl D, Rogers S, Adhya S, Merril CR. Bacteriophage K1-5 encodes two different tail fiber proteins, allowing it to infect and replicate on both K1 and K5 strains of Escherichia coli. JVirol. 2001;75:2509–15. doi:10.1128/JVI.75.6.2509-2515.2001.
11. Ahmed R, Bopp C, Borczyk A, Kasatiya S. Phage-typing scheme for Escherichia coli O157:H7. J Infect Dis. 1987;155:806–9.
12. Weitz JS, Poisot T, Meyer JR, Flores CO, Valverde S, Sullivan MB, et al. Phage-bacteria infection networks. Trends Microbiol. 2013;21:82–91. doi:S0966-842X (12) 00200-4;10.1016/j.tim.2012.11.003.
13. Hausmann R. Bacteriophage T7 genetics. Curr Top Microbiol Immunol. 1976;75:77–110.
14. Miller ES, Kutter E, Mosig G, Arisaka F, Kunisawa T, Ruger W. Bacteriophage T4 genome. MicrobiolMolBiol Rev. 2003;67:86–156.
15. Kropinski AM, Lingohr EJ, Moyles DM, Chibeu A, Mazzocco A, Franklin K, et al. Escherichia coli O157:H7 typing phage V7 is a T4-like virus. JVirol. 2012;86:10246. doi:86/18/10246;10.1128/JVI.01642-12.
16. Kropinski AM, Waddell T, Meng J, Franklin K, Ackermann HW, Ahmed R, et al. The host-range, genomics and proteomics of Escherichia coli O157:H7 bacteriophage rV5. VirolJ. 2013;10:76. doi:1743-422X-10-76;10.1186/1743-422X-10-76.
17. Perry LL, SanMiguel P, Minocha U, Terekhov AI, Shroyer ML, Farris LA, et al. Sequence analysis of Escherichia coli O157:H7 bacteriophage PhiV10 and identification of a phage-encoded immunity protein that modifies the O157 antigen. FEMS MicrobiolLett. 2009;292:182–6. doi:FML1511;10.1111/j.1574-6968.2009.01511.x.
18. Santos MA. An improved method for the small scale preparation of bacteriophage DNA based on phage precipitation by zinc chloride. Nucleic Acids Res. 1991;19:5442.
19. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18:821–9. doi:gr.074492.107;10.1101/gr.074492.107.
20. Seemann, T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014. doi:10.1093/bioinformatics/btu153First published online: March 18, 2014.
21. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. JComputBiol. 2012;19:455–77. doi:10.1089/cmb.2012.0021.
22. McCune B, Grace JB. Analysis of ecological communities. Gleneden Beach, Ore: MjM Software; 2002.
23. Liu X, Murata T. An efficient algortihm for optimizing bipartite modularity in bipartite networks. JACIII. 2010;14:408–15.
24. Beckett, SJ. A weighted modularity algorithm for bipartite networks. figshare. 2014. http://dx.doi.org/10.6084/m9.figshare.999114.
25. Beckett SJ, Boulton CA, Williams HT. FALCON: a software package for analysis of nestedness in bipartite networks. F1000Res. 2014;3:185. doi:10.12688/f1000research.4831.1.
26. Almeida-Neto M, Guimaraes P, Guimaraes Jr PR, Loyola RD, Ulrich W. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. Oikos. 2008;117:1227–39.
27. Atmar W, Patterson BD. The measure of order and disorder in the distribution of species in fragmented habitat. Oecologia. 1993;96:373–82.
28. Oksanen, J, Blanchet, FG, Kindt, R, Legendre, P, Minchin, PR, O'Hara, RB, Simpson, GL, Solymos, P, Stevens, MHH, Wagner H. vegan: community ecology package. R package version 2.0-10. 2013. http://CRAN.R-project.org/package=vegan.
29. Brualdi RA, Sanderson JG. Nested species subsets, gaps, and discrepancy. Oecologia. 1999;119:256–64.
30. Bascompte J, Patterson BD. The nested assembly of plant-animal mutualistic networks. Proc Natl Acad Sci U S A. 1993;100:9383–0387.
31. Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMCGenomics. 2011;12:402. doi:1471-2164-12-402;10.1186/1471-2164-12-402.
32. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. Bioinformatics. 2011;27:1009–10. doi:btr039;10.1093/bioinformatics/btr039.
33. Matthews L, Reeve R, Gally DL, Low JC, Woolhouse ME, McAteer SP, et al. Predicting the public health benefit of vaccinating cattle against Escherichia coli O157. Proc Natl Acad Sci U S A. 2013;110:16265–70. doi:1304978110;10.1073/pnas.1304978110.
34. Leiman PG, Battisti AJ, Bowman VD, Stummeyer K, Muhlenhoff M, Gerardy-Schahn R, et al. The structures of bacteriophages K1E and K1-5

explain processive degradation of polysaccharide capsules and evolution of new host specificities. JMolBiol. 2007;371:836–49. doi:S0022-2836 (07) 00756-5;10.1016/j.jmb.2007.05.083.

35. Hattman S, Fukasawa T. Host-induced modification of t-even phages due to defective glucosylation of their DNA. Proc Natl Acad Sci U S A. 1963;50:297–300.

36. Springman R, Badgett MR, Molineux IJ, Bull JJ. Gene order constrains adaptation in bacteriophage T7. Virology. 2005;341:141–52. doi:S0042-6822 (05) 00418-6;10.1016/j.virol.2005.07.008.

37. Beckett SJ, Williams HT. Coevolutionary diversification creates nested-modular structure in phage-bacteria interaction networks. Interface Focus. 2013;3:20130033. doi:10.1098/rsfs.2013.0033; rsfs20130033.

38. Flores CO, Meyer JR, Valverde S, Farr L, Weitz JS. Statistical structure of host-phage interactions. Proc Natl Acad Sci U S A. 2011;108:E288–97. doi:1101595108;10.1073/pnas.1101595108.

39. Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasic L, Thingstad TF, Rohwer F, et al. Explaining microbial population genomics through phage predation. NatRevMicrobiol. 2009;7:828–36. doi:nrmicro2235;10.1038/nrmicro2235.

40. Mora A, Blanco M, Blanco JE, Alonso MP, Dhabi G, Thomson-Carter F, et al. Phage types and genotypes of shiga toxin-producing Escherichia coli O157:H7 isolates from humans and animals in spain: identification and characterization of two predominating phage types (PT2 and PT8). JClinMicrobiol. 2004;42:4007–15. doi:10.1128/JCM.42.9.4007-4015.2004;42/9/4007.

41. Baggesen DL, Sorensen G, Nielsen EM, Wegener HC. Phage typing of Salmonella Typhimurium - is it still a useful tool for surveillance and outbreak investigation? Euro Surveill. 2010;15:19471.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

MICROBIAL GENOMICS
Bases to Biology

Research Paper

Short-term evolution of Shiga toxin-producing *Escherichia coli* O157:H7 between two food-borne outbreaks

Lauren A. Cowley,[1] Timothy J. Dallman,[1] Stephen Fitzgerald,[2] Neil Irvine,[3] Paul J. Rooney,[4] Sean P. McAteer,[2] Martin Day,[1] Neil T. Perry,[1] James L. Bono,[5] Claire Jenkins[1] and David L. Gally[2]

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

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

[3]Public Health Agency, 12-22 Linenhall St, BT2 8BS Belfast, Northern Ireland

[4]Microbiology Laboratory, Royal Victoria Hospital, BT12 6BA Belfast, Northern Ireland

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

Correspondence: Lauren A. Cowley (lauren.cowley@phe.gov.uk)

DOI: 10.1099/mgen.0.000084

Shiga toxin-producing *Escherichia coli* (STEC) O157:H7 is a public health threat and outbreaks occur worldwide. Here, we investigate genomic differences between related STEC O157:H7 that caused two outbreaks, eight weeks apart, at the same restaurant. Short-read genome sequencing divided the outbreak strains into two sub-clusters separated by only three single-nucleotide polymorphisms in the core genome while traditional typing identified them as separate phage types, PT8 and PT54. Isolates did not cluster with local strains but with those associated with foreign travel to the Middle East/North Africa. Combined long-read sequencing approaches and optical mapping revealed that the two outbreak strains had undergone significant microevolution in the accessory genome with prophage gain, loss and recombination. In addition, the PT54 sub-type had acquired a 240 kbp multi-drug resistance (MDR) IncHI2 plasmid responsible for the phage type switch. A PT54 isolate had a general fitness advantage over a PT8 isolate in rich medium, including an increased capacity to use specific amino acids and dipeptides as a nitrogen source. The second outbreak was considerably larger and there were multiple secondary cases indicative of effective human-to-human transmission. We speculate that MDR plasmid acquisition and prophage changes have adapted the PT54 strain for human infection and transmission. Our study shows the added insights provided by combining whole-genome sequencing approaches for outbreak investigations.

Keywords: Escherichia coli; Bioinformatics; Evolution; Prophage; Recombination.

Abbreviations: PT, phage type; TP, typing phage; STEC, shiga toxin-producing *Escherichia coli*; SNP, single-nucleotide polymorphism; WGS, whole-genome sequencing.

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

Data Summary

1. Short read FASTQ sequences have been deposited in the NCBI Short Read Archive under the BioProject PRJNA248042 (http://www.ncbi.nlm.nih.gov/bioproject/248042).
2. Long read FASTA files are deposited in NCBI Genbank under accessions CP015831 (644-PT8 chromosome) (http://www.ncbi.nlm.nih.gov/nuccore/Cp015831), CP015832 (180-PT54 chromosome) (http://www.ncbi.nlm.nih.gov/nuccore/CP015832) and CP015833(180-

Received 9 June 2016; Accepted 08 August 2016

Downloaded from www.microbiologyresearch.org by
IP: 134.174.250.136
On: Sat, 03 Dec 2016 18:09:30

© 2016 The Authors . Published by Microbiology Society
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/).

PT54 plasmid) (http://www.ncbi.nlm.nih.gov/nuccore/CP015833).

Introduction

Shiga toxin-producing *E. coli* (STEC) serogroup O157:H7 can cause severe bloody diarrhoea and haemolytic uraemic syndrome and has been a significant public health threat since it emerged in 1982 (Pennington, 2010). Several key virulence factors contribute to pathogenicity; these include the production of Shiga toxin (Stx) and expression of a type three secretion system (Law, 2000). STEC O157:H7 is a globally disseminated pathogen and emerged approximately 120 years ago (Dallman *et al.*, 2015a).

STEC O157:H7 is a zoonotic pathogen and transmission is commonly associated with direct or indirect contact with animals, especially ruminants, their environment, or the consumption of contaminated food or water. Foodborne outbreaks have been linked to fast-food outlets and restaurants (Bell *et al.*, 1994). Historically, outbreaks were detected and investigated by comparing phage type, pulsed-field gel electrophoresis patterns or multilocus variable number tandem repeat analysis (MLVA) (Byrne *et al.*, 2014). Phage typing has been used for surveillance and outbreak investigations at the Gastrointestinal Bacteria Reference Unit (GBRU) in the United Kingdom since 1992 (Khakhria *et al.*, 1990) and provides a low-cost and rapid test to broadly discriminate between strains. More recently, whole-genome sequencing (WGS) has been used to facilitate STEC O157:H7 outbreak investigations, and strains exhibiting less than five single-nucleotide polymorphisms (SNPs) in the core genome are likely to be temporally linked and share a common source (Dallman *et al.*, 2015b). High-resolution genome-based methods have been use to track human infections phylogenetically (Eppinger *et al.*, 2011; Jenkins *et al.*, 2015).

The sequencing of the Sakai and EDL933 genomes (Hayashi *et al.*, 2001; Latif *et al.*, 2014) showed that both genomes contained an array of integrated prophages with variation of the prophage content between the two strains. Prophages containing the Shiga toxin-encoding genes (*stx*) are known to exhibit variation between STEC O157 strains (Allison, 2007; Eppinger *et al.*, 2011; Herold *et al.*, 2004; Ogura *et al.*, 2015). Little is known however about the short-term micro-evolution of STEC O157:H7 genomes, for example during outbreaks. In part, this is due to difficulties with the assembly of repetitive and paralogous features of prophages when using short-read sequencing caused by multiple assignment in the genomes when reads do not span repeated paralogous prophage genes. Long-read sequencing technologies, such as PacBio or MinION, have been shown to achieve improved *de novo* assemblies that facilitate more accurate characterization of the accessory genome (Cooper *et al.*, 2014; Latif *et al.*, 2014) including prophage regions (Asadulghani *et al.*, 2009).

This study describes a public health investigation of two related outbreaks of STEC O157:H7 associated with the same food outlet in 2012 for which an outbreak report has recently been released (http://www.publichealth.hscni.net/publications/report-outbreak-control-team-investigations-outbreak-e-coli-o157-associated-flicks-rest). In August 2012, four cases of STEC O157:H7 phage type (PT) 8 were epidemiologically-linked to the consumption of food at this restaurant. Eight weeks later, in October 2012, over 140 confirmed cases (and >160 unconfirmed cases) of STEC O157:H7 PT54 were also linked to the same restaurant, with 15 confirmed cases being the result of secondary transmission. MLVA profiles from both incidents indicated that the August and October outbreaks were caused by the same strain despite the phage typing difference. Both short- and long-read sequencing was used to characterize the micro-evolutionary events that occurred in the core and accessory genome between the first (PT8) and the second (PT54) clusters of cases. Our research focused on isolate variation between the two related outbreaks in order to try and understand the much larger scale of the second (PT54) outbreak. The work has facilitated significant insights into short-term changes that can occur in the STEC O157:H7 genome associated with human infection and provides important lessons for outbreak investigations involving this zoonotic pathogen.

Impact Statement

In this article, we explore the changes in the genome of a strain of STEC O157:H7 that caused two food-borne outbreaks associated with the same restaurant that were only 8 weeks apart. Utilising sequence data from three different sequencing platforms we provide evidence of short-term evolution between strains isolated in the two outbreaks. This included multi-drug resistance plasmid acquisition and phage content variation, including duplication, that has occurred since the outbreak isolates diverged from a common ancestor over an estimated 1-year period. Based on growth and competitive index assays, we speculate that the genomic changes may account for the higher number of cases associated with the second outbreak. This work has highlighted the value of combining different sequencing and *in vitro* approaches to assist investigations into the epidemiology of outbreaks.

Methods

PT and MLVA analysis. All 145 cultures, received at the reference laboratory, from cases linked to both the August and October clusters were typed by phage typing and MLVA (Byrne *et al.*, 2014; Khakhria *et al.*, 1990). The MLVA profiles for the outbreak strains isolated in August and October were mostly 5-8-12-4-5-2-8-3. However, the profile sshowed a high degree of variation at VNTR locus #3 but all isolates were a single-locus variant (SLV) of each other. Isolates that have the same MLVA profile or SLV of that profile are regarded as microbiologically linked (Byrne

Downloaded from www.microbiologyresearch.org by
IP: 134.174.250.136
On: Sat, 03 Dec 2016 18:09:30

et al., 2014). MVLA analysis had revealed that isolates from both the PT54 and PT8 outbreaks clustered together and that, despite their different PTs, the two occurrences of STEC O157:H7 at the food outlet were likely to be associated with very closely related strains.

Illumina sequencing and core SNP analysis. As part of the outbreak investigation 89 isolates were selected for WGS, including four from the August PT8 cluster and 53 from the PT54 cluster in October and 30 isolates of STEC O157:H7 from temporally and geographically related sporadic cases isolated between June and November 2012. Genomic DNA was fragmented and tagged for multiplexing with Nextera XT DNA Sample Preparation Kits (Illumina) and sequenced at the Animal Laboratories and Plant Health Agency using the Illumina GAII platform with 2×150 bp reads. Short reads were quality trimmed (Bolger *et al.*, 2014) and mapped to the reference STEC O157 strain Sakai (Genbank accession BA000007) using BWA-SW (Li & Durbin, 2009). The sequence alignment map output from BWA was sorted and indexed to produce a binary alignment map (BAM) using Samtools (Li & Durbin, 2009). GATK2 (McKenna *et al.*, 2010) was used to create a variant call format (VCF) file from each of the BAMs, which were further parsed to extract only SNP positions which were of high quality [mapping quality (MQ)>30, depth (DP)>10, genotype quality (GQ)>30, variant ratio >0.9]. Pseudosequences of polymorphic positions were used to reconstruct maximum-likelihood trees using RaxML (Stamatakis, 2014). Pair-wise SNP distances between each pseudosequence were calculated. Spades version 2.5.1 (Bankevich *et al.*, 2012) was run using careful mode with kmer sizes 21, 33, 55 and 77 to produce *de novo* assemblies of the sequenced paired-end fastq files. FASTQ sequences were deposited in the NCBI Short Read Archive under the BioProject PRJNA248042.

PacBio sequencing. One isolate of STEC O157 PT8 (ref 644-PT8) and one belonging to PT54 (ref 180-PT54) were selected. High-molecular-weight DNA was extracted using Qiagen Genomic-tip 100/G columns and a modification of the protocol previously described by Clawson *et al.* (2009). Samples (10 µg) of DNA was sheared to a targeted size of 20 kb using a g-TUBE (Corvaris) and concentrated using 0.45×volume of AMPure PB magnetic beads (Pacific Biosciences) following the manufacture's protocol. Sequencing libraries were created using 5 µg of sheared DNA and the PacBio DNA SMRTbell Template Prep Kit 1.0 and fragments 10 kb or larger selected using a BluePippin (Sage Science) with the smrtbell 15–20 kb setting. The library was bound with polymerase P5 followed by sequencing on a RS II sequencing platform (Pacific Biosciences) with chemistry C3 and the 120 min data collection protocol.

A fastq file was generated from the sequencing reads using SMRTanalysis and error-corrected reads were created using PBcR with self-correction (Koren *et al.*, 2013). The longest 20× coverage of the corrected reads were assembled with Celera Assembler 8.1. The resulting contigs were polished using Quiver (Chin *et al.*, 2013) and annoted using PROKKA (Seemann, 2014). The annotated genome sequence was imported into Geneious (Biomatters) and duplicated sequence removed from the 5′ and 3′ ends to generate the circularized chromosome. The origin of replication was approximated using OriFinder (Luo *et al.*, 2014) and the chromosome reoriented using the origin as base 1. PacBio sequenced strain 180-PT54 is available under accession numbers CP015832 for the chromosome and CP015833 for the plasmid.

MinION sequencing. DNA from 644-PT8 was extracted using the STRATEC molecular invisorb spin minikit and diluted to a concentration of 1 µg of genomic DNA in 50 µl of water. The MinION library was prepared using the SQK-MAP006 genomic sequencing kit according to the manufacturer's instructions and sequencing was performed on a Mk1 MinION with a Mk1 flow cell.

Approximately 76 000 reads were produced and 26-fold passing 2D coverage of the genome was achieved. The long-read assembly program Canu version 1.1 (Koren *et al.*, 2013) was used to assemble the long reads and two chromosomal contigs were produced. The assembly of 644-PT8 showed greater concordance with the synteny of 180-PT54 than the assembly of the PacBio sequencing had produced, but was still not resolved in a similar region.

OpGen mapping for the isolate was obtained from a commercial provider. 644-PT8 was rotated using OriFinder to have the same point of origin as isolate 180-PT54 for comparison. The genome was then annotated using PROKKA (Seemann, 2014). MinION sequenced strain 644-PT8 is available under the accession number CP015831.

Analysis of the accessory genome using long-read sequences. The two annotated chromosome assemblies were analysed in PHAST (Zhou *et al.*, 2011) which identifies complete and incomplete prophage regions and their constituent genes.

The gene annotations and their sequences for each strain were compared with each other with blastn (Altschul *et al.*, 1990). Unique genes were extracted from the results if they had no match at greater than 70 % nucleotide identity and overlap to any of the annotated genes in the other strain using a reciprocal blast approach. Roary (Page *et al.*, 2015) was used to confirm whether the identified genes were representative of the rest of the outbreak, i.e. that PT54-specific genes were missing from all PT8 isolates and PT8-specific genes were missing from all PT54 isolates.

NUCMER (Kurtz *et al.*, 2004) was used to align the two assemblies and to identify SNPs, 23 012 ambiguous alignment SNPs were excluded. These SNPs were confirmed by aligning the Illumina-sequenced contigs also to confirm that the SNPs were found with both short -and long-read techonologies.

Gene annotations were extracted from the genbank file and resistance annotations were manually analysed for differences between the two strains. Plasmid differences were

Downloaded from www.microbiologyresearch.org by
IP: 134.174.250.136
On: Sat, 03 Dec 2016 18:09:30

visualised in Mauve (Darling *et al.*, 2010). The plasmid from isolate 180-PT54 was graphically visualised using BRIG (Alikhan *et al.*, 2011) and compared with other IncHI2 plasmid sequences found in Genbank that were highly similar by blast at >98 % identity and >60 % coverage (accession numbers KM877269.1, JN983042.1, BX664015.1, DQ517526.1, EF382672.1, LN794248.1, LK056646.1, EU855787.1, KP975077.1, EU855788.1, CP011601.1, CP008906.1, CP008825.1, CP012170.1 and CP006056.1).

Plasmid conjugation. Nalidixic-acid-resistant (NalR) colonies of 644-PT8 were isolated from overnight cultures on LB-agar with 20 µg nalidixic acid ml^{-1}. Conjugation was performed on LB-agar by co-streaking donor 180-PT54 and recipient spontaneous NalR of 644-PT8. Co-streaked growth was harvested in phosphate-buffered saline then plated onto LB-agar with 20 µg nalidixic acid ml^{-1} and 10 µg chloramphenicol ml^{-1}. Resistant colonies were purified by streaking onto fresh plates of 20 µg nalidixic acid ml^{-1} and 10 µg chloramphenicol ml^{-1}.

Acid-resistance assays. Acid-resistance assays were performed as described previously (Castanie-Cornet *et al.*, 1999). Briefly, cells were cultured overnight in either LBG [Luria-Bertani (LB) broth + 0.4 % glucose], LB buffered with 100 mM morpholinepropanesulfonic acid (MOPS pH8) or LB buffered with 100 mM morpholineethanesulfonic acid (MES pH 5.5). Overnight (22 h) stationary-phase cultures were diluted 1:1000 into pre-warmed minimal E glucose (EG) media, pH 2.5. The glutamate- and arginine-dependent systems were tested by growing cells overnight in LBG and diluting cultures into EG (pH 2.5) supplemented with either 1.5 mM glutamate or 0.6 mM arginine, respectively. The glucose-repressed system was tested by growing cells overnight in LB-MES pH 5.5 followed by dilution in EG pH 2.5. Overnight cultures grown in either LB-MOPS (pH 8) or LBG followed by dilution in unsupplemented EG were used as acid-sensitive controls for the glucose-repressed and glutamate- or arginine-dependent AR systems, respectively. Viable cells were enumerated at t=0 and t=4 h and used to calculate percentage survival.

Fitness assays. Fitness of 180-PT54 relative to 644-PT8 was calculated as described previously (Lenski, 1991). Viable-cell counts for each competing strain were determined at time zero (t=0) and again after 24 h of co-culturing by selective plating. Fitness was calculated using the formula:

Fitness index (f.i.) = LN (N_i (1)/ N_i (0)) / LN (N_j (1)/ N_j (0)),

Where N_i (0) and N_i (1) = initial and final colony counts of strain 180-PT54, respectively and

N_j (0) and N_j (1) = initial and final colony counts of strain 644-PT8, respectively (Lenski, 1991).

Biolog phenotyping microarray. A single isolated Shiga toxin-containing *Escherichia coli* O157:H7 colony was grown on BUG+B agar overnight at 33 °C. A sterile swab was used to transfer cells from the plate into inoculating fluid 0 (IF-0) to a turbidity of 43 % T (transmittance) and addition IF-0 with dye was to a final cell density of 85 % T. For phenotyping microarray (PM) plates 1 and 2 (Biolog), 100 µl per well was added. PM plates 3, 4, 6, 7 and 8 were supplemented with 20 mM sodium succinate and 2 µM ferric citrate before 100 µl was added to each well (Bochner *et al.*, 2001). All plates were incubated at 33 °C for 48 h using the Omnilog II Combo System (Biolog). The output data from the Omnilog was imported into the opm package in R for analysis (Vaas *et al.*, 2013).

Results

Short-read sequencing analysis demonstrates that the outbreak strains are closely related and were not endemic

Analysis of Illumina sequences indicated that all four isolates from the August PT8 outbreak had identical core genome sequences. There were three SNP differences in the core genome between the PT8 isolates from August and the PT54 isolates from October. The maximum distance between isolates within the October PT54 cluster was four SNPs, including acquisition of a maximum of two SNPs from a common haplotype. Previous temporal analysis of STEC WGS data predicts a mutation rate of approximately 2.5 SNPs per year, therefore it was likely their last common ancestor was very recent (approximately 1 year) but prior to the occurrence of the two public health incidents (Dallman *et al.*, 2015a). The phylogeny of the outbreak isolates indicated that the PT54 cluster did not directly evolve from the PT8 cluster but instead that they share a very recent common ancestor. Furthermore, it was evident that the PT8 and PT54 strains were closely related to each other but genetically distinct from strains of STEC O157:H7 circulating in the local population (Figs 1 and 2). Strains held in the Public Health England (PHE) STEC O157 WGS database that clustered most closely with the outbreak strains were associated with foreign travel to Egypt and Israel (Fig. 2). Although the precise source was never identified by the investigation that followed it was likely that the strains were imported in contaminated food with the larger outbreak possibly exacerbated by an infected or colonised food handler in the restaurant.

Prophage variation identified between specific PT8 and PT54 isolates based on combined long-read sequencing approaches

PacBio sequencing enabled assemblies of the genome of 180-PT54 into one contig and the genome of 644-PT8 into two contigs. Difficulties with the assembly of 644-PT8 were a result of the reorganized synteny of the genome compared with the closely related PT54 isolate. From comparison of the MinION assembly and the OpGen map, it was clear that the disrupted assembly was caused by a 200 kbp inverted repeat in the genome that constituted the second smaller

Downloaded from www.microbiologyresearch.org by
IP: 134.174.250.136
On: Sat, 03 Dec 2016 18:09:30

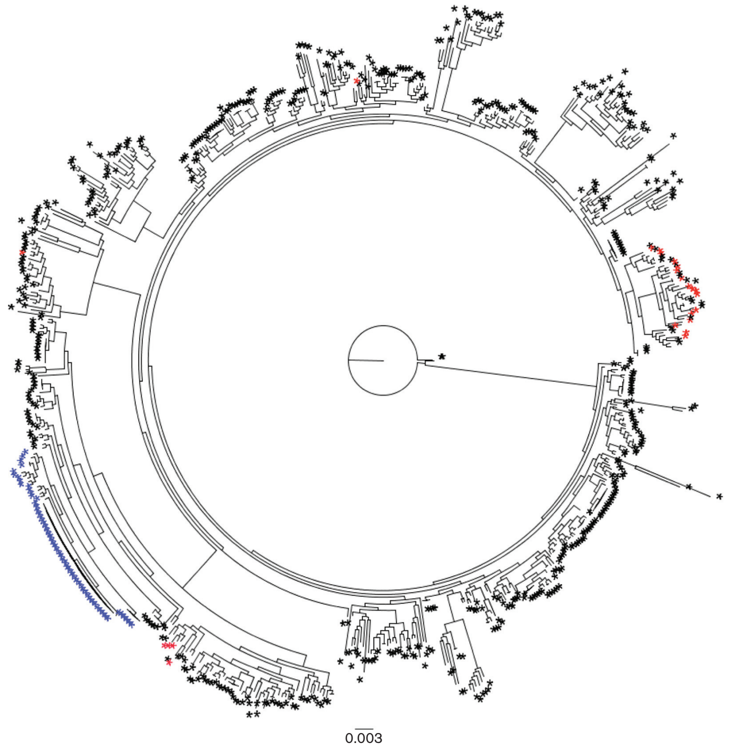


Fig. 1. Maximum-likelihood phylogenetic tree of STEC O157 strains selected for various PTs in PHE database associated with domestically acquired infection and travel-related cases. SNPs called on core genome via a mapping technique against the reference strain Sakai. Outbreak strains are indicated in blue and local background strains in the region where the outbreak occurred are in red.

contig in the assembly. The contig was inserted twice into the larger contig aligning with the NCoI sites found in the OpGen map (Fig. S1, available in the online Supplementary Material). The combination of the MinION sequencing assembly and the OpGen map enabled us to construct a single contig of 5.8 Mb for the isolate that included the 200 kbp repeat.

A set of 14 prophage regions was shared between the representative isolates of the PT8 and PT54 clusters. In addition to the 14 shared prophage regions, 180-PT54 had gained one prophage region of 24 874 bp located at 2 281 433–2 306 307 bp and 644-PT8 had acquired one prophage region of 20 818 bp located at 4 773 172–4 793 990 bp. However, the subsequent Roary analysis showed that the PT8 unique prophage was not missing from all the PT54 outbreak isolates so was not specific to the PT8 outbreak. The 180-PT54 unique prophage was likely to be specific to the PT54 outbreak as it was missing from all the PT8 isolates. In addition, the genome of 644-PT8 had three repeated prophages within the 200 kbp inverted repeat. Two shared prophage regions, designated P7 and P8 were similar prophages that showed variation between the two representative isolates, indicative of recombination that had contributed to the inverted repeat (Fig. S2). The prophage changes between long-read sequenced isolate 644-PT8 and 180-PT54 are detailed in Fig. 3 and those changes that are representative of the rest of that PT sub-cluster are indicated by asterisks.

All the unique gene differences within the chromosome of each representative strain are listed in Table S1. Those that were confirmed by Roary to be representative of the rest of that PT outbreak are highlighted in bold type. The genes that vary between 180-PT54 and 644-PT8 are detailed in

Downloaded from www.microbiologyresearch.org by
IP: 134.174.250.136
On: Sat, 03 Dec 2016 18:09:30

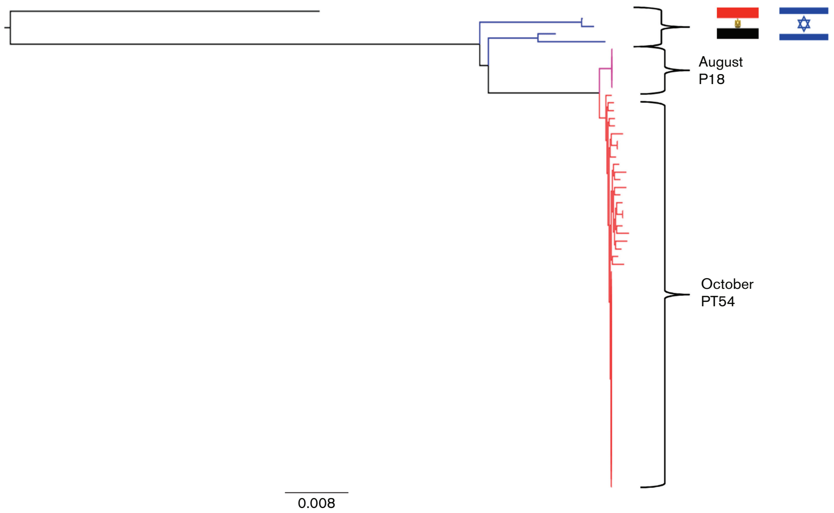


Fig. 2. Maximum-likelihood phylogenetic tree of outbreak strains and most closely related strains in the PHE database that are associated with foreign travel to Egypt and Israel. The blue branches represent strains that are associated with travel to Egypt and Israel, the pink branches represent the original PT8 outbreak and the red branches represent the later PT54 outbreak.

Fig. 3. 644-PT8 had 82 unique genes on the chromosome, 79 of which were prophage-associated but the Roary analysis showed that these changes were only present in a subset of the PT54 isolates and not all members of the rest of the PT54 outbreak. 180-PT54 had 30 unique genes, 25 of which were prophage-associated and these were found to be missing from all the other PT8 isolates in the outbreak and therefore were representative of the PT54 outbreak.

Whole-genome alignment using NUCMER identified 29 SNPs between the two isolates (644-PT8 and 180-PT54) in fully aligned regions. The SNP locations and base changes are detailed in Table 1 and have been confirmed in the Illumina data. This was higher than the three SNPs identified between the 'core' genomes based on the short-read sequencing but 26 of these SNPs were found in the P7 and P8 shared prophage regions. These prophage regions would not have been shared with Sakai so would not have been called in the original core genome SNP-calling from the Illumina data.

Plasmid acquisition by 180-PT54

Both strains harboured pO157, the O157 virulence plasmid present in nearly all strains of STEC O157:H7 (Lim *et al.*, 2010). However, isolate 180-PT54 acquired an additional approximately 220 genes introduced on an IncHI2 plasmid (Johnson *et al.*, 2006) not present in 644-PT8. While both 180-PT54 and 644-PT8 exhibited tellurite and tetracycline resistance, 180-PT54 was also resistant to chloramphenicol and streptomycin and this matched with resistance genes located on the IncHI2 plasmid. The IncHI2 plasmid was predicted to encode at least six membrane proteins, a drug efflux pump, other resistance mechanisms including additional tellurite resistance and protection from exposure to heavy metal ions (mercury). It also encoded at least two DNA methylases (Fig. 4). Relatedness depicted in a BRIG plot (Fig. 4) demonstrates the high similarity with other IncHI2 plasmids that have been detected in clinical isolates worldwide (Chen *et al.*, 2007; Feasey *et al.*, 2014; Gilmour *et al.*, 2004; Kariuki *et al.*, 2015; Li *et al.*, 2013).

Phage type transition is associated with plasmid acquisition

It was suggested that the IncHI2 plasmid may be responsible for the difference in PT observed between the two outbreak clusters in August and October. To test this, the IncHI2 plasmid was conjugated into 644-PT8 and the conjugant was then phage typed. Acquisition of the plasmid, as defined by inheritance of chloramphenicol resistance, converted 644-PT8 to PT54. Antibiotic resistance and replicon type profiling using the Illumina Roary data demonstrated that the plasmid-associated resistance was present in all the PT54 isolates in this study and that they all carried the IncHI2 plasmid associated with the PT transition. Analysis of the genes present on the plasmid (Fig. 4) shows a number of determinants that could be associated with changes in phage resistance including

Downloaded from www.microbiologyresearch.org by
IP: 134.174.250.136
On: Sat, 03 Dec 2016 18:09:30

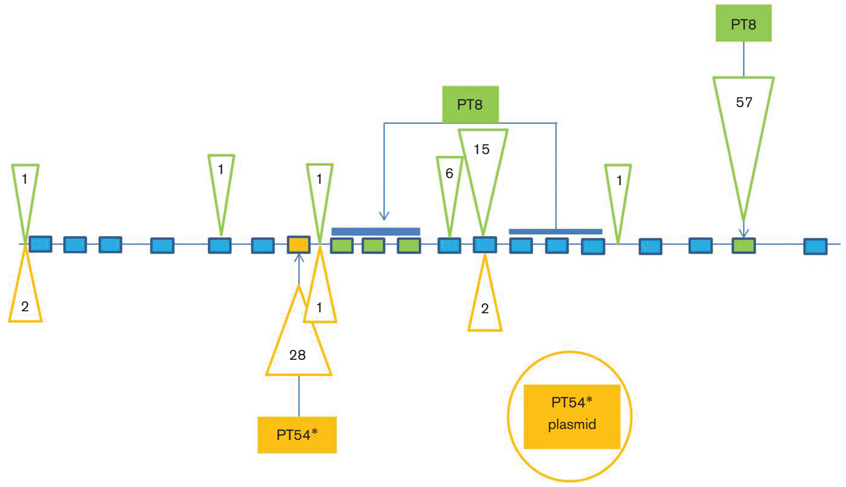


Fig. 3. Schematic diagram representing accessory genome variation between the two long-read-sequenced outbreak isolates (180-PT54 and 644-PT8). Blue rectangles represent shared prophage regions, the orange rectangle represents a unique 180-PT54 prophage region and green rectangles represent unique 644-PT8 prophage regions. Orange triangles represent locations and number of unique genes for 180-PT54. The inverted repeat region is indicated by a blue line above the repeated prophage blocks. The unique plasmid in the PT54 outbreak is represented by the orange circle. Those changes that have been confirmed to be representative of all the other members of that PT sub-cluster by Roary analysis have an asterisk next to them.

tellurite resistance (Whelan *et al.*, 1997). While both 644-PT8 and 180-PT54 have chromosomally-encoded tellurite resistance, specifically *terW*, only 180-PT54 have *terY* and *terX* as these are present on the IncHI2 plasmid. Methylase-modification genes encoded on the plasmid are also potential candidates to confer resistance to specific bacteriophages (Labrie *et al.*, 2010).

Increased fitness of the PT54 strain associated with the larger second outbreak

Acquisition of IncHI2 plasmids commonly confers phage, antibiotic and heavy metal resistance, thus increasing bacterial fitness under certain environmental conditions (Fang *et al.*, 2016; Whelan *et al.*, 1995). Conversely an increased metabolic burden imposed by the 200 kbp inversion during DNA replication in strain 644-PT8 is likely to reduce fitness. To assess any differences in fitness, strains 644-PT8 and 180-PT54 were competed by co-culturing in LB-broth at 37 °C and 25 °C. Under both conditions 180-PT54 significantly outcompeted 644-PT8 (fitness index=1.28 and 1.23, respectively). Fitness was therefore independent of culture temperature. Subsequently Biolog phenotypic microarrays were performed to determine the nature of the observed fitness increase. Similar growth was observed for both strains for the majority of carbon, phosphorus and sulphur sources tested (data not shown) however growth of 180-PT54 was increased for multiple nitrogen sources, including certain amino acids and dipeptides (Fig. S3). This is in agreement with the observed increased fitness of 180-PT54 when cultured in LB-broth in which amino acids/short peptides are the primary carbon and nitrogen sources. In addition, growth of 180-PT54 was better than that of 644-PT8 when ammonia was the sole nitrogen source (Fig. S3). Of the multiple di- and tri-peptide nitrogen sources on which 180-PT54 grew better than 644-PT8 many contained arginine or glutamate. *E. coli* possesses three acid-resistance systems (AR) of which two are dependent on arginine and glutamate, respectively (Richard & Foster, 2003). We therefore tested if AR was altered in 180-PT54 by the increased metabolism of arginine and glutamate relative to 644-PT8. For each AR system (glucose-repressed, arginine- and glutamate-dependent) strain 644-PT8 was significantly more resistant to acid shock when either pre-adapted in LB pH 5.5 or supplied with exogenous Arg or Glu (Table 2). Without pre-adaptation however strain 180-PT54 was more acid-resistant than 644-PT8 (P=0.035).

Discussion

Phylogeny techniques based on 'core genome' sequence analyses have been transformative for epidemiological investigations and also provide an assessment of evolutionary relationships between strains (Holt *et al.*, 2012; Quick *et al.*, 2015). This study focused on two temporally related

Downloaded from www.microbiologyresearch.org by
IP: 134.174.250.136
On: Sat, 03 Dec 2016 18:09:30

Table 1. Table listing the positions and base changes of all the SNPs found between the PT8 strain and the PT54 strain in a whole-genome alignment using the program NUCMER

The SNPs identified by the previous phylogenetic analysis using Illumina data are highlighted in bold and italicized type, the third SNP identified by the phylogenetic analysis is within a repeat region so was excluded as ambiguous alignment by the program NUCMER. All other identified SNPSs, not in bold, are part of the mobilome.

PT54 position	PT54 base	PT8 base	PT8 position
1975308	G	C	1974495
2681706	C	A	3282472
2681715	T	C	3282463
2681722	T	C	3282456
2681730	T	G	3282448
2681757	T	G	3282421
2681766	G	T	3282412
2681775	A	G	3282403
2681784	C	T	3282394
2681788	G	A	3282390
2681796	T	C	3282382
2681823	G	A	3282355
2681833	G	A	3282345
2681835	C	T	3282343
2681844	A	G	3282334
2681847	G	A	3282331
2681865	G	A	3282313
2681973	C	T	3282205
2681976	C	G	3282202
2681977	T	C	3282201
2681979	T	A	3282199
2681981	A	C	3282197
2681982	C	A	3282196
2681985	C	A	3282193
2833323	***A***	***C***	***3027880***
2894348	***A***	***G***	***3088905***
3048043	C	A	3242600
3048059	C	A	3242616
3048069	G	A	3242626

outbreaks of STEC O157:H7 from the same restaurant. Initially, MLVA and phage typing results were contradictory as MLVA indicated that the outbreaks were caused by the same strain although the phage types were distinct. The relatedness of the PT8 and PT54 strains was confirmed by short-read sequencing which defined three SNP differences in the core genome between the two groups of strains indicating that they share a very recent common ancestor (approximately 1 year). Switching of PT within a sublineage has been observed previously (Dallman *et al.*, 2015b) but this is some of the first, to our knowledge, documented evidence of PT conversion within two closely related outbreaks and the mechanisms behind that.

In this study, the application of PacBio and MinION sequencing, as well as OpGen mapping, enabled us to obtain single-contig assemblies of two isolates associated with the two outbreaks at the restaurant, one belonging to PT8 and one belonging to PT54. These assemblies clearly showed the high prophage carriage in these isolates, which is typical of STEC O157:H7 with approximately 12–14 % of the genome made up of highly paralogous phage genes (Fig. S2). Analysis of the genomes from the long-read sequencing showed that there had been a shift in prophage composition between the two outbreak groups. There was gain and loss of prophage while two of the shared prophage regions had undergone recent recombination. Furthermore, isolate 644-PT8 had a repeat of three of the shared prophage regions in a 200 kbp inverted region. There is significant, apparent functional redundancy across the different prophages. This is in agreement with earlier research from Hayashi and colleagues that showed a diverse bacteriophage complement produced from a single strain including recombination between prophage loci (Asadulghani *et al.*, 2009).

Isolate 180-PT54 assembled into three contigs; the chromosome, the F-like pO157 and an IncHI2 plasmid. The IncHI2 plasmid was large (240 kbp) and predicted to encode about 220 genes. This plasmid was not present in 644-PT8. IncHI2 plasmids are commonly associated with the spread of extended-spectrum β-lactam resistance (ESBL) genes, heavy metal resistance and phage resistance (Whelan *et al.*, 1995; Fang *et al.*, 2016). We identified several antibiotic and environmental resistance genes potentially conferring resistance to chloramphenicol, streptomycin, tellurite, tetracycline and certain heavy metal ions encoded on the incHI2 plasmid acquired by 180-PT54 (Fig. 4). The resistance loci facilitated conjugation of the plasmid into 644-PT8, leading to the recipient strain phage typing as PT54. Conversion of PT8 to PT54 is caused by the acquisition of resistance to the group 3 typing phages (TP4, TP5 and TP14) (Cowley *et al.*, 2015). There is a previous report that an IncHI2 plasmid can confer bacteriophage resistance (Whelan *et al.*, 1997) therefore adding to the possible survival advantage conferred on strains that acquire this plasmid. A BLAST search revealed that highly similar IncHI2 plasmids have been described in several different organisms from around the world (Fig. 4), including Taiwan, China, Kenya, Malawi and the USA. Identification of incHI2 plasmids in *E. coli* however is rare (Fang *et al.*, 2016; Losada *et al.*, 2016). The acquisition of this plasmid is therefore likely to increase the survival capacity of the strain under certain stressful environmental conditions.

In addition to antibiotic and phage resistance, we demonstrated that 180-PT54 was significantly fitter than 644-PT8 under a defined set of growth conditions. The genetic differences identified in 180-PT54, that include plasmid acquisition, resulted in a fundamental alteration in central nitrogen metabolism that enhanced growth compared with 644-PT8. In accordance with the trade-off between self-preservation and nutritional competency (SPANC) the increased growth of 180-PT54 also resulted in decreased acid resistance, at least under priming conditions (Ferenci,

Downloaded from www.microbiologyresearch.org by
IP: 134.174.250.136
On: Sat, 03 Dec 2016 18:09:30

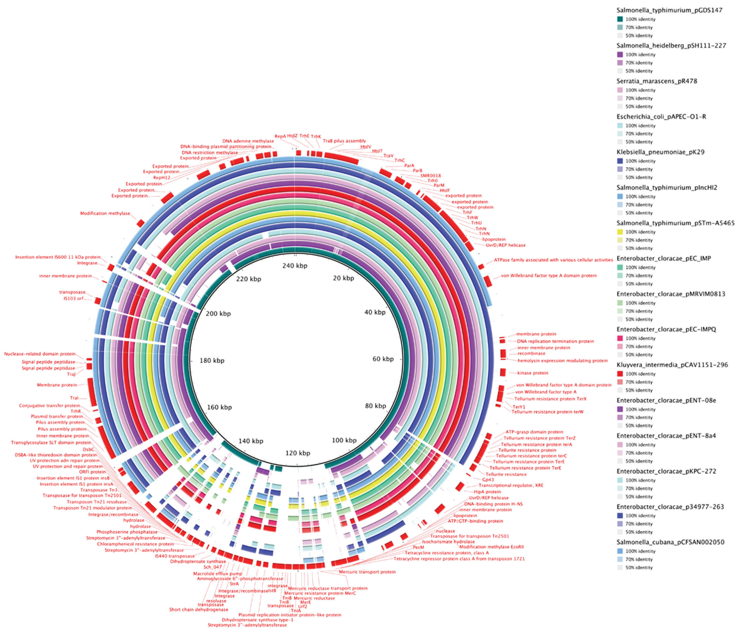


Fig. 4. BRIG plot of the approximately 240 kbp IncHI2 plasmid found in 180-PT54 as the central reference showing genomic similarity between it and other IncHI2 plasmids found in Genbank from various species of bacteria. Annotations are shown in red on the outermost ring. The darker the colour the greater the level of genomic similarity between the 180-PT54 IncHI2 plasmid and the plasmid found in a different organism. The plasmids are indicated by the colours labelled vertically at the right hand side of the figure.

2005). The public health investigation of the outbreaks included sampling of employees who worked at the restaurant. Two members of staff were shown to be colonized with the PT54 strain during the outbreak and an analysis identified a significant association with one of these employees and the infection risk (http://www.publichealth.hscni.net/publications/report-outbreak-control-team-investigations-outbreak-e-coli-o157-associated-flicks-rest), although it is not known if this individual could have actually accounted for the second outbreak. Based on this and the altered genotype and phenotype of the PT54 strain, we speculate that it may be more adapted for human colonization than the original PT8 strain and that this capacity may be linked to the much higher number of cases associated with the second outbreak. These included multiple examples of human-to-human transmission. STEC O157:H7 strains are usually associated with ruminant hosts and presumably human colonization could lead to adaptive changes that promote survival, such as plasmid acquisition and prophage variation. Sequencing of human cases in the UK has identified a subset of imported strains that are

Table 2. Table describing the results of the acid-resistance assays performed on strains 644-PT8 and 180-PT54 to assess their biological fitness

Adaptation medium*	Challenge medium (pH2.5)	Percentage survival†	
		PT54	PT8
LB pH8	EG	0.811582004	0.187176
LB pH 5.5	EG	4.09726344	38.78565
LBG	EG	14.77276286	7.230324
	EG+Glu	21.71965529	52.69993
	EG+Arg	37.2807674	70.11365

*Strains were adapted by overnight culturing in either LB pH 5.5 or LBG before diluting 1:1000 in EG pH 2.5 or supplemented EG. †The percentage survival was determined after 4 h of acid challenge. The mean percentage survival of six replicates (n=6) is shown for EG challenge and three replicates (n=3) for EG supplemented with either glutamate (1.5 mM) or arginine (0.6 mM).

Downloaded from www.microbiologyresearch.org by
IP: 134.174.250.136
On: Sat, 03 Dec 2016 18:09:30

acquired during travel abroad or brought in by colonised foreign visitors (Dallman *et al.*, 2015a). These strains are significantly more likely to contain plasmids encoding antibiotic resistance and there is a concern that acquisition of such elements may adapt strains to enable transmission and/or persistence in the human population, which would be a serious public health concern. However, we do acknowledge that there are other possible reasons why a greater number of cases were associated with the second outbreak; these include the possibility of different contaminated products in the second outbreak or increased awareness that might introduce a bias in recorded cases.

This outbreak investigation illustrated the power of both short- and long-read sequencing technologies to investigate and understand foodborne outbreaks. The evolutionary context illustrated by the short-read WGS data revealed the true genetic relationship between the strains from the August and October clusters and provided evidence of the geographical origin of the strains. The geographical signal derived from WGS data will greatly facilitate outbreak investigation where imported food is implicated. The use of long-read WGS data clearly demonstrated the dynamic nature of the accessory genome in STEC O157:H7 and the potential impact of horizontal gene transfer over a short time frame. The long-read sequencing enabled us to identify a plasmid that confers resistance to antibiotics and bacteriophages as well as other environmental stressors. The plasmid was shown to causes phage-type conversion in a strain of STEC O157:H7.

Acknowledgements

We would like to acknowledge the outbreak control team and their valuable epidemiological investigation into this outbreak, especially:

Louise Herron (Public Health Agency, Northern Ireland), Damian Connolly (Belfast City Council) and Maria T Jennings (Food Standards Agency, Northern Ireland).

References

Alikhan, N. F., Petty, N. K., Ben Zakour, N. L. & Beatson, S. A. (2011). BLAST ring image generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* **12**, 402.

Allison, H. E. (2007). Stx-phages: drivers and mediators of the evolution of STEC and STEC-like pathogens. *Future Microbiol* **2**, 165–174.

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.

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.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S. & other authors (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**, 455–477.

Bell, B. P., Goldoft, M., Griffin, P. M., Davis, M. A., Gordon, D. C., Tarr, P. I., Bartleson, C. A., Lewis, J. H., Barrett, T. J. & Wells, J. G. (1994). A multistate outbreak of *Escherichia coli* O157:H7-associated bloody diarrhea and hemolytic uremic syndrome from hamburgers. The Washington experience. *JAMA* **272**, 1349–1353.

Bochner, B. R., Gadzinski, P. & Panomitros, E. (2001). Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Genome Res* **11**, 1246–1255.

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

Byrne, L., Elson, R., Dallman, T. J., Perry, N., Ashton, P., Wain, J., Adak, G. K., Grant, K. A. & Jenkins, C. (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**, e85901.

Castanie-Cornet, M. P., Penfound, T. A., Smith, D., Elliott, J. F. & Foster, J. W. (1999). Control of acid resistance in *Escherichia coli. J Bacteriol* **181**, 3525–3535.

Chen, Y. T., Lauderdale, T. L., Liao, T. L., Shiau, Y. R., Shu, H. Y., Wu, K. M., Yan, J. J., Su, I. J. & Tsai, S. F. (2007). Sequencing and comparative genomic analysis of pK29, a 269-kilobase conjugative plasmid encoding CMY-8 and CTX-M-3 β-lactamases in *Klebsiella pneumoniae. Antimicrob Agents Chemother* **51**, 3004–3007.

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.

Cooper, K. K., Mandrell, R. E., Louie, J. W., Korlach, J., Clark, T. A., Parker, C. T., Huynh, S., Chain, P. S., Ahmed, S. & Carter, M. Q. (2014). Complete genome sequences of two *Escherichia coli* O145:H28 outbreak strains of food origin. *Genome Announc* **2**, e00482–14.

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**.

Dallman, T. J., Ashton, P. M., Byrne, L., Perry, N. T., Petrovska, L., Ellis, R., Allison, L., Hanson, M. F., Holmes, A. & other authors (2015a). Applying phylogenomics to understand the emergence of Shiga-toxin-producing *Escherichia coli* O157:H7 strains causing severe human disease in the UK. *Microbial Genomics* **1**.

Dallman, T. J., Byrne, L., Ashton, P. M., Cowley, L. A., Perry, N. T., Adak, G., Petrovska, L., Ellis, R. J., Elson, R. & other authors (2015b). Whole-genome sequencing for national surveillance of Shiga toxin-producing *Escherichia coli* O157. *Clin Infect Dis* **61**, 305–312.

Darling, A. E., Mau, B. & Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* **5**, e11147.

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.

Fang, L., Li, X., Li, L., Li, S., Liao, X., Sun, J. & Liu, Y. (2016). Co-spread of metal and antibiotic resistance within ST3-IncHI2 plasmids from *E. coli* isolates of food-producing animals. *Sci Rep* **6**, 25312.

Feasey, N. A., Cain, A. K., Msefula, C. L., Pickard, D., Alaerts, M., Aslett, M., Everett, D. B., Allain, T. J., Dougan, G. & other authors (2014). Drug resistance in *Salmonella enterica* ser. Typhimurium bloodstream infection, Malawi. *Emerg Infect Dis* **20**, 1957–1959.

Downloaded from www.microbiologyresearch.org by
IP: 134.174.250.136
On: Sat, 03 Dec 2016 18:09:30

Ferenci, T. (2005). Maintaining a healthy SPANC balance through regulatory and mutational adaptation. *Mol Microbiol* **57**, 1–8.

Gilmour, M. W., Thomson, N. R., Sanders, M., Parkhill, J. & Taylor, D. E. (2004). The complete nucleotide sequence of the resistance plasmid R478: defining the backbone components of incompatibility group H conjugative plasmids through comparative genomics. *Plasmid* **52**, 182–202.

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.

Herold, S., Karch, H. & Schmidt, H. (2004). Shiga toxin-encoding bacteriophages-genomes in motion. *Int J Med Microbiol* **294**, 115–121.

Holt, K. E., Baker, S., Weill, F. X., Holmes, E. C., Kitchen, A., Yu, J., Sangal, V., Brown, D. J., Coia, J. E. & other authors (2012). *Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat Genet* **44**, 1056–1059.

Jenkins, C., Dallman, T. J., Launders, N., Willis, C., Byrne, L., Jorgensen, F., Eppinger, M., Adak, G. K., Aird, H. & other authors (2015). Public health investigation of two outbreaks of Shiga toxin-producing *Escherichia coli* O157 associated with consumption of watercress. *Appl Environ Microbiol* **81**, 3946–3952.

Johnson, T. J., Wannemeuhler, Y. M., Scaccianoce, J. A., Johnson, S. J. & Nolan, L. K. (2006). Complete DNA sequence, comparative genomics, and prevalence of an IncHI2 plasmid occurring among extraintestinal pathogenic *Escherichia coli* isolates. *Antimicrob Agents Chemother* **50**, 3929–3933.

Kariuki, S., Okoro, C., Kiiru, J., Njoroge, S., Omuse, G., Langridge, G., Kingsley, R. A., Dougan, G. & Revathi, G. (2015). Ceftriaxone-resistant *Salmonella enterica* serotype typhimurium sequence type 313 from Kenyan patients is associated with the *bla*$_{\mathrm{CTX\text{-}M15}}$ gene on a novel IncHI2 plasmid. *Antimicrob Agents Chemother* **59**, 3133–3139.

Khakhria, R., Duck, D. & Lior, H. (1990). Extended phage-typing scheme for *Escherichia coli* O157:H7. *Epidemiol Infect* **105**, 511–520.

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.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C. & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12.

Labrie, S. J., Samson, J. E. & Moineau, S. (2010). Bacteriophage resistance mechanisms. *Nat Rev Microbiol* **8**, 317–327.

Latif, H., Li, H. J., Charusanti, P., Palsson, B. Ø. & Aziz, R. K. (2014). A gapless, unambiguous genome sequence of the enterohemorrhagic *Escherichia coli* O157:H7 strain EDL933. *Genome Announc* **2**.

Law, D. (2000). Virulence factors of *Escherichia coli* O157 and other Shiga toxin-producing *E. coli*. *J Appl Microbiol* **88**, 729–745.

Lenski, R. E. (1991). Quantifying fitness and gene stability in microorganisms. *Biotechnology* **15**, 173–192.

Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760.

Li, L., Liao, X., Yang, Y., Sun, J., Li, L., Liu, B., Yang, S., Ma, J., Li, X. & other authors (2013). Spread of *oqxAB* in *Salmonella enterica* serotype Typhimurium predominantly by IncHI2 plasmids. *J Antimicrob Chemother* **68**, 2263–2268.

Lim, J. Y., La, H. J., Sheng, H., Forney, L. J. & Hovde, C. J. (2010). Influence of plasmid pO157 on *Escherichia coli* O157:H7 Sakai biofilm formation. *Appl Environ Microbiol* **76**, 963–966.

Losada, L., DebRoy, C., Radune, D., Kim, M., Sanka, R., Brinkac, L., Kariyawasam, S., Shelton, D., Fratamico, P. M. & other authors (2016). Whole genome sequencing of diverse Shiga toxin-producing and non-producing *Escherichia coli* strains reveals a variety of virulence and novel antibiotic resistance plasmids. *Plasmid* **83**, 8–11.

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.

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.

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.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., Fookes, M., Falush, D., Keane, J. A. & Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693.

Pennington, H. (2010). *Escherichia coli* O157. *Lancet* **376**, 1428–1435.

Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., Nair, S., Neal, K., Nye, K. & other authors (2015). Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. *Genome Biol* **16**, 114.

Richard, H. T. & Foster, J. W. (2003). Acid resistance in *Escherichia coli*. *Adv Appl Microbiol* **52**, 167–186.

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

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

Vaas, L. A., Sikorski, J., Hofner, B., Fiebig, A., Buddruhs, N., Klenk, H. P. & Göker, M. (2013). opm: an R package for analysing OmniLog® phenotype microarray data. *Bioinformatics* **29**, 1823–1824.

Whelan, K. F., Colleran, E. & Taylor, D. E. (1995). Phage inhibition, colicin resistance, and tellurite resistance are encoded by a single cluster of genes on the IncHI2 plasmid R478. *J Bacteriol* **177**, 5016–5027.

Whelan, K. F., Sherburne, R. K. & Taylor, D. E. (1997). Characterization of a region of the IncHI2 plasmid R478 which protects *Escherichia coli* from toxic effects specified by components of the tellurite, phage, and colicin resistance cluster. *J Bacteriol* **179**, 63–71.

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

1. Dallman, T. J., Ashton, P. A., Jenkins, C., Grant K. NCBI Short Read Archive PRJNA248042 (2015).
2. Cowley, L. A., Dallman, T. J., Bono, J. NCBI Genbank CP015831, CP015832 and CP015833 (2015).

Downloaded from www.microbiologyresearch.org by
IP: 134.174.250.136
On: Sat, 03 Dec 2016 18:09:30