UNIVERSITY OF
LEICESTER

Genomics and population dynamics of phase variable genes in *Campylobacter*

Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

by

Jack Aidley

Department of Genetics
University of Leicester
Leicester, UK

May 2017

In memory of

David John Aidley

1937 - 2000

Rest in peace, Dad

Abstract

Phase variation is a feature of many pathogenic bacteria, including *Campylobacter jejuni* – a leading cause of food-borne gastroenteritis. *C. jejuni* has many phase variable (PV) genes which exhibit high frequency, stochastic, reversible switching between ON and OFF expression states. This results from instability in simple sequence repeats (SSRs).

This study is the first to conduct a widescale survey of the *Campylobacter* 'phasome' (the set of PV genes present in a genome). A new program, Phasome*It*, was developed to identify and compare all SSR-mediated phase variable genes in 77 complete *Campylobacter* genomes. Surprisingly, there are a large number of rare PV genes with few, or no, homologues in other genomes. These exist alongside a "core phasome" of PV genes associated with particular species. This suggests a significant role for phasome diversity in *Campylobacter* population and disease dynamics.

SSR tract length influences rates of PV, however it is not known whether differences in PV rate are biologically significant or whether a threshold effect exists. A cyclical assay was developed which allows alternation of experimental conditions favouring the ON and OFF states of *cj1421c*. Bacteriophage F336 selects for the OFF state whilst human serum selects for the ON state, and both produce complete selective sweeps. These agents were combined for a complete ON→OFF→ON cycle. Using an *in silico* model of this assay it was demonstrated that a broad range of conditions favour PV over non-PV genes whilst variation rate is primarily dependent on duration of selection. Extending this modelling approach to the interaction of non-selective bottlenecks and PV loci indicated that smaller bottlenecks have a disruptive effect on population dynamics whilst larger bottlenecks preserve increased diversity. These differences are capable of producing stochastic differences in output populations that may drive host-to-host diversity and result in rare occurrence of disease sequelae.

Contents

List of Abbreviations

AAA+	ATPases Associated with diverse cellular Activities
ATP	adenosine triphosphate
BHI	brain-heart infusion
BLAST	Basic Local Alignment Search Tool
bp	base pair(s)
BSA	bovine serum albumin
CBHI	cation-enriched brain-heart infusion
CFU	colony forming units
CPS	capsular polysaccharides
CSV	comma separated values
DNA	2-deoxyribonucleic acid
EOP	efficiency of plating
FASTA	FAST-All
FSA	Food Standards Agency
GBS	Guillain Barré syndrome
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HTML	hypertext markup language
IBS	irritable bowel syndrome
IQR	interquartile range
LA	lysogeny broth agar
LB	lysogeny broth
LOS	lipooligosaccharide
LPS	lipopolysaccharide
MHA	Mueller-Hinton agar
MHB	Mueller-Hinton broth

MIT	Massachusetts Institute of Technology
MLST	multi-locus sequence typing
MOI	multiplicity of infection
mRNA	messenger RNA
NCBI	National Center for Biotechnology Information
NEB	New England BioLabs Inc.
NZCYM	NZ-amine, casamino acids, yeast extract, magnesium
OD	optical density
ORF	open reading frame
PCR	polymerase chain reaction
PFU	plaque forming units
PNACL	The Protein and Nucleic Acids Chemistry Lab
PV	phase variable
RCF	relative centrifugal force
RNA	ribonucleic acid
rRNA	ribosomal RNA
SAM	S-adenosyl-L-methionine
SM	saline magnesium
SSR	simple sequence repeat
ST	sequence type
TFB1	transformation buffer 1
TFB2	transformation buffer 2
tRNA	transfer RNA

List of Figures

List of Tables

Acknowledgements

As with any significant piece of work, this thesis was not achieved alone but with the help of a great many others. First and foremost among those is my supervisor, Dr. Christopher D. Bayliss, without whose wisdom, friendship and patient guidance none of this would have been possible. I have received help from a great number of the current and past members of Lab 121 but, in particular, Dr. Richard Haigh has taught me a huge number of techniques and answered a countless number of my questions. I'd also like to give particular thanks to Dr. Alex Woodacre, Dr. Rachael Madison, Dr. Amelia Veselis, Dr. Luke Crane, Shane Hussey, and Neelam Dave. As part of my project I travelled to Copenhagen where I worked in the lab of Dr. Lone Brønsted under the guidance of Dr. Martine Holst Sørensen and I would like to thank them both for their patience, knowledge and guidance. I would also like to thank Professor Samuel K. Shepherd at the University of Bath, both for access to his genome collection and for his help and assistance in the phasome project, and to Dr. Guillaume Méric for preparing these genomes in GenBank format. I am also grateful for the guidance provided by my thesis committee, Professors Ed Louis and Julian Ketley.

Many thanks to Dr. Daniela Rudloff, Dr. Helen Collins, Dr. Luke Green, Joe Wanford, and, of course, Chris for proof reading this thesis in its various stages. Any remaining mistakes are my own.

Outside of the lab, I am extremely grateful for the love and support of my partner, Daniela, particularly during the final few months of thesis writing and also for the love and support of my friends and family. My parents raised me with love and affection and, both scientists themselves, gave me an early love of science and the natural world that has carried me to where I am today. Throughout my PhD I was funded by the BBSRC through the MIBTP programme and I'd like to extend my thanks to Dr. Peter Meacock, Dr. Ezio Rosato, Professor Chris Thomas, Professor Richard Napier, and Dr. Katherine Denby for their efforts in organising the MIBTP. Finally, I'd like to thank the Street fund, the Bond fund and – especially – the Microbiology Society for the funding they provided for conference travel and for my research visit to Copenhagen.

This thesis was typeset in LaTeX, using MikTeX 2.9.

Chapter 1

Introduction

1.1 Overview

Bacteria face a diverse range of challenges and have evolved a complex and varied repertoire of strategies for dealing with these challenges. One of these strategies is phase variation, which is a form of stochastic phenotype switching involving random switches between a defined number of states in a heritable manner. This process produces a high degree of population heterogeneity in a short period of time and thus allows the bacterium to adapt to a changing environment at the population level. Phase variation is present in a diverse range of bacteria, including *Campylobacter jejuni* which is the leading cause of food poisoning in the developed world.

Phase variation is involved in a range of significant bacterial traits, including motility, phage resistance, and immune evasion, some of which may impact on virulence. Understanding the principles behind the evolution and behaviour of

phase variation has a significant role to play in our understanding of the spread and pathogenicity of these organisms. *C. jejuni* has a large number of phase variable genes which operate by a simple mechanism and at particularly high frequencies, thus this organism is amenable for use as a model organism for studying questions related to phase variation.

1.2 Phase variation is high-frequency, reversible change in phenotype

Phase variation refers to high frequency, stochastic, reversible, and heritable change in phenotype (van der Woude and Bäumler, 2004). Although "high frequency" is a subjective term, phase variation is usually employed to cover mutations occurring roughly every 10^2 to 10^5 divisions and confined to microbial species (Moxon *et al.*, 2006) although it has been applied to mutation rates as high as 1 in every 10 divisions (van der Woude and Bäumler, 2004). Phase variation is most commonly the result of hypermutable genomic regions that influence phenotype, but epigenetic switches may also give rise to the phenomena.

1.2.1 Historical view

Phase variation was first reported in 1922 by Andrewes who observed reversible switching in a strain of *Salmonella enterica* (referred to at the time as *Bacillus paratyphosus*) responsible for paratyphus B infection. In this work, he was investigating the specific and non-specific binding of serum raised against a strain. He discovered that there appeared to be a clear divide into two distinct groups in terms of their response to serum, and that isolates raised from both groups could spontaneously switch to the other group (Andrewes, 1922). This kind of switching

behaviour was rapidly confirmed by other researchers in similar organisms (Topley and Ayrton, 1924, Bensted, 1925). Although Andrewes (1922) does not employ the term 'phase variation' or speak of the two states being 'phases', his subsequent paper (Andrewes, 1925) refers to the 'phase' of 'diphasic salmonellas'. The term 'phase variation' was introduced into the literature by Kauffman and Mitsui (1930) (rendered in the German as 'Phasenwechsel') for a similar phenomenon of 'alpha-beta phase variation' also discovered in *Salmonella*. Further work into the phenomena discovered by Andrewes steadily unravelled the mechanism of this action (e.g. Stocker, 1949, Lederberg and Iino, 1956) until it was finally determined that it operated by the means of a genetic inversion between two genes H1 and H2 which controlled a surface antigen on the flagella (Zieg *et al.*, 1977). In the intervening fifty years, reports of similar phenomena in a range of bacteria were identified (e.g. Veazie, 1949, Stoker and Fiset, 1956, Nakase *et al.*, 1969), although particular interest was paid to the more narrow form of antigenic variation in which key antigens involved in immune recognition and response are subject to stochastic variation.

1.2.2 Prevalence

Phase variation is widespread in prokaryotes, having seemingly independently evolved in multiple taxa (Orsi *et al.*, 2010, Ackermann and Chao, 2006, Mrázek *et al.*, 2007, Lin and Kussell, 2012). It occurs in a wide range of pathogenic bacteria – e.g. *Bacillus anthracis*, *Bacteroides fragilis*, *Campylobacter jejuni*, *Mycobacterium tuberculosis*, *Neisseria meningitidis*, *Streptococcus pneumoniae*, and *Yersinia pestis* (Cerdeno-Tarraga *et al.*, 2005, Lin and Kussell, 2012, Manso *et al.*, 2014) – but also in plant pathogens (e.g. *Burkholderia cenocepacia*), insect symbionts (e.g. *Buchnera aphidicola*), free-living cyanobacteria (e.g. *Prochlorococcus marinus*) and in archaea – e.g. *Methanococcus maripaludis* and *Sulfolobus islandicus* (Lin and Kussell, 2012). Prevalence across this wide range of environments suggests that phase variation is a broadly useful

strategy rather than just having a narrow range of benefits. In many cases, these organisms contain multiple phase variable traits which can give rise in a combinatorial fashion to a broad range of possible phenotypes (Henderson *et al.*, 1999, Bayliss *et al.*, 2001, van der Woude and Bäumler, 2004).

Single celled eukaryotes also employ forms of stochastic phenotype switching. Perhaps the most widely studied is mating type switching in which yeast spontaneously alter their mating type[1] allowing clonal populations derived from asexual division from a single cell to undergo meiosis (Klar, 2010, Haber, 2012). Even among multi-cellular organisms analogous mechanisms are used to provide diversity and mosaic effects in the somatic cells, most importantly in the generation of diversity in the immune system (Tonegawa, 1983, Rajewsky, 1996), but also in generating phenotypic diversity in endothelial cells (Yuan *et al.*, 2016).

1.2.3 Common mechanisms

Phase variation can be mediated by a variety of genetic mechanisms (reviewed in van der Woude and Bäumler, 2004), such as inversion of DNA sections (e.g. Cerdeno-Tarraga *et al.*, 2005, Manso *et al.*, 2014), or insertion-excision mechanisms (e.g. Perkins-Balding *et al.*, 1999, Loessner *et al.*, 2002), however the primary mechanism in *Campylobacter* is instability in simple sequence repeats. This mechanism is used in a variety of bacterial species and operates via instability in simple sequence repeats in which a short motif (1-10bp) is precisely repeated one after the other. These repeats are unstable during replication resulting in gain or loss of a single repeat unit with high frequency by insertion or deletion, likely due to slipped-strand mispairing (Levinson and Gutman, 1987, Henderson *et al.*, 1999). Slipped-strand mispairing occurs during replication where there is a separation

[1]Mating type is analogous to sex in that only yeast of different mating types can mate but lacks the distinct phenotypic profiles of true sexual systems.

of the template and copy strands followed by a re-annealing in which the strands fail to correctly align (i.e. there is slippage). This slippage produces a kink in either the template or copy strand that effectively hides a single repeat unit from the replication machinery. If this kink occurs on the copy strand then there is an insertion event as an extra repeat unit is replicated, whereas if it occurs on the template strand there is a deletion event as one less repeat unit is replicated. This is illustrated in figure 1.1. Note that because this process influences only the newly copied strand the mutation is only passed onto one of the two daughter cells.

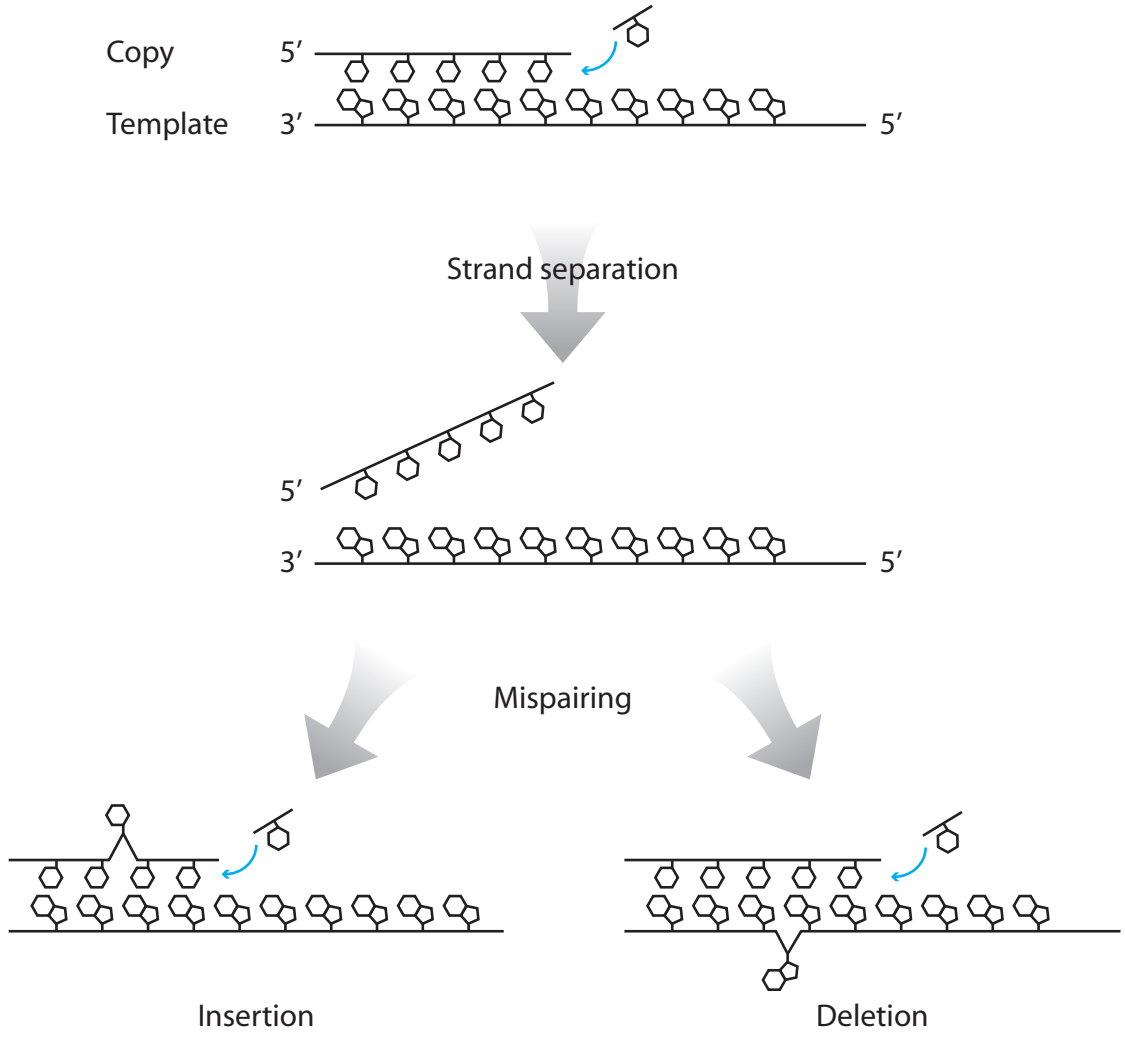

Figure 1.1: Slipped strand mispairing can produce either insertion or deletion
This diagram illustrates possible slipped strand mispairing events. During replication, the template and copy strands separate, on re-annealing there can be a kink in either the copy (shown on left) or template (shown on right) strand resulting in an insertion (left) or deletion (right) when replication resumes.

Changes in length through slipped-strand mispairing are reversible because an

insertion can be reversed by a deletion and *vice-versa*. These length changes can produce changes in phenotype by altering gene expression in one of two primary ways: (1) if located within the reading frame of a gene it can produce a frameshift mutation resulting in a truncated, non-functional protein; or (2) if located in the promoter region it can produce a change in gene expression levels by altering the distance between promoter elements (van Ham *et al.*, 1993, Moxon *et al.*, 2006, Alamro *et al.*, 2014). In *C. jejuni*, the known repeats producing phase variation are all poly-G/C tracts of between 7 and 13 base pairs in length (Parkhill *et al.*, 2000).

1.2.4 Scoring and phasotype

As alluded to above, there are frequently multiple phase variable traits in any individual organism, each of which can independently adopt different phase states. The combination of phase states in an organism is referred to as its 'phasotype' (Alamro *et al.*, 2014). The phasotype is distinct from the genotype because it is based on scoring the state of multiple phase variable loci and the same scored state can be produced by multiple possible genotypes (e.g. if more than one simple sequence repeat (SSR) length produces the ON state). It is distinct from the phenotype for two reasons: (1) because it is not necessarily known how gene expression influences phenotype so the state is based on a putative change in expression rather than a phenotypic response, and (2) because it is possible for the state of one gene to effect the impact of another. For example, in *Neisseria gonorrhoeae* the α side change genes have a combinatorial effect in which the first gene, *lgtA* adds a group which is then further modified by the second, *lgtD*, so if the first is not active the second has no effect in either state (Yang and C., 1996).

Because these combinations of states are combinatorial in effect, the number of possible phasotypes an organism can exist in increases rapidly with the number of phase variable loci. Under the assumption of two phase states per loci this leads

to 2^n possible states for n loci, or an approximate thousand-fold increase in the number of states per ten loci. This means that 29 loci would lead to over half a billion possible states, and 24 loci to close to 17 million, while a more modest six loci would lead to 64 possible phasotypes.

1.2.5 Biological and evolutionary relevance

Since its discovery in 1922, the functional and fitness benefits of phase variation have been subject to much speculation. Because there is a large diversity in the function of identified phase variable genes, it is unlikely that a single explanation suffices for all genes. In general the possible drivers are (1) adaptation to a changing environment, (2) frequency selection, and (3) re-establishment of phenotypic diversity after a population bottleneck. In the first of these, switching may allow the organism to persist in environments which vary in an alternating fashion (Beaumont *et al.*, 2009, Libby and Rainey, 2011), possible examples of this include changes in agglutination and motility (Karlyshev *et al.*, 2002, van Alphen *et al.*, 2008), or opposing states of a single phase variable gene aiding cell invasion and serum resistance (van Alphen *et al.*, 2014). In frequency selection, neither state is fully favoured but rather selection favours a mixed population, this may be the case with restriction modification (R/M) systems which may prevent the accumulation of resistant (methylated) phage within the population (Bayliss *et al.*, 2006). Variation in these systems may also influence the expression of other genes producing a "phasevarion" of genes that alter their state together (Srikhanta *et al.*, 2005; 2010; 2011, Manso *et al.*, 2014). Finally, phenotypic variation within populations may assist in immune evasion (van der Woude and Bäumler, 2004, Alamro *et al.*, 2014) or provide population-level protection against phage which are dependent upon certain surface structures to bind to the cell (Sørensen *et al.*, 2011) and phase variation allows the rapid re-establishment of varied phenotypes within a population following population bottlenecks.

1.2.6 Computer modelling of phase variation

The use of mathematics to predict and understand the behaviour of biological systems has a long history, with models such as the Lotka-Volterra equation (Lotka, 1910; 1920) being among famous early examples. Technological developments have allowed models to become more complex and permitted them to be studied in more detail compared to the slow progress of manual calculation. These computational approaches have come to be referred to as *in silico* experiments.

Phase variation is amenable to *in silico* approaches as it is believed to operate under simple probabilistic mechanics. These mechanics are easily translated into mathematical models and computer simulations. The earliest relevant work considered simply the ideal mutation rate between two states under conditions that alternated selection for these two states. Their analytical approach yielded an ideal mutation rate, μ_{ess}, of $\frac{1}{T}$ where T is the number of generations the environment is stable for (Leigh, 1970). However, an alternative model, published in Gillespie (1981) which studied mutation rates in a randomly and rapidly fluctuating environment found that the ideal mutation rate in asexual organisms was always zero under these circumstances. Gillespie attributed the difference between their model and that reported in Leigh (1970) to differences in the rate of fluctuation modelled.

Although these general models are applicable to the specific case of phase variation, they were aimed instead at understanding the drivers behind mutation rates in general. More recent models have directly focused on understanding phase variation in particular. Salathé *et al*. (2009) used a stochastic model where the population could move between one of two phenotypic states and the mutation rate between states could be directional so that mutation from the first state into the second could be more or less likely than the reverse mutation from the second state into the first. They looked at both symmetrical and asymmetrical selective forces. Their

work indicated that phase variation could only be maintained under asymmetrical selection if the selection co-efficient was high. Libby and Rainey (2011)'s model, created to explain the results of an experimental evolution experiment carried out by Beaumont *et al.* (2009) which demonstrated that stochastic phenotype switching could be induced *in vitro*[2], indicated that strong 'exclusion rules' (i.e. selective bottlenecks) could favour phase variation under a broad range of conditions.

Palmer *et al.* (2013) went further by using a model which modelled not only two selection states but also the changing rate of mutation in an underlying model of phase variation. To study this they used a model based on the phase variable loci of *Haemophilus influenzae* which modelled the changing length of the tracts as well as the changes in mutation rate associated with tract length. In *H. influenzae*, the rate of variation of the phase variable *mod* gene (which codes for a type III DNA methyltransferase) had previously been shown to vary linearly with the length of the repeat tract (which has a 5′-AGTC repeat unit) from 17 to 38 repeats and showed a 2:1 bias for deletions over insertions (De Bolle *et al.*, 2000). The model used this data to simulate changes in tract length with the tract allowed to vary between 11 and 61 repeats. Selection was then applied as a simple probability of survival based on whether the number of repeats coded for a putative ON state (i.e. if it represented a multiple-of-3 change in length and thus returned the frame to that in the 11-repeat ON state) and the selection alternated between selecting for ON and selecting for OFF. They investigated selection under both symmetric and asymmetric conditions and concluded that selection for phase variation would only occur when the selection co-efficient or the period of stability in both directions were sufficient to completely select for the favoured phenotype. They expressed this rule with an observed critical value as $sT > 7$ where s is the selection co-efficient (defined so that the relative fitness of the state selected against is $1 - s$) and T is the period of stability before switching, although conditions of symmetric

[2] Although the experimental result of Beaumont *et al.* (2009) turned out to be a form of bistability (Gallie *et al.*, 2015) rather than phase variation *per se* the model developed by Libby and Rainey (2011) is directly relevant to phase variation.

selection permit phase variation to be favoured at lower sT values.

1.3 *C. jejuni* is an important human pathogen

1.3.1 A historical perspective on *Campylobacter* discovery

The first species from the genus now known as *Campylobacter* was isolated in 1906 by John McFadyean who was seeking to understand the causes of an epidemic of foetal abortion in sheep and isolated distinctive spiral bacteria from many of the aborted foetuses (Skirrow, 2006). An apparently identical species was identified later the same year by Smith and Taylor as involved in foetal abortion in cattle, and they gave the species the name *Vibrio fetus* in a later publication (Smith and Taylor, 1919). Although occasional reports of human infection occur in the literature stretching back to 1913 (e.g. Curtis, 1913, Ward, 1948) it was regarded as an occasional scientific curio until the 1950s when more serious investigation was carried out by Elizabeth King who isolated similar-looking bacteria from diarrhoeal human stool samples which she dubbed 'related vibrios' (King, 1957). Sporadic reports of further cases were published through the 60s and early 70s (e.g. Middelkamp and Wolf, 1961, Darrell *et al.*, 1967, McDonald and Mautner, 1970, Dekeyser *et al.*, 1972) until Butzler *et al.* conducted a more systematic search for 'related vibrios' (Butzler *et al.*, 1973). Butzler *et al.* were able to isolate *Campylobacter* spp. from just over 5% of the 800 cases they investigated. In the same year it was recognised that the two species most common in human disease – *C. jejuni* and *Campylobacter coli* – should be included into the genus *Campylobacter* (Véron and Chatelain, 1973) which one of the same authors had coined as the new designation for *Vibrio fetus* ten years earlier (Sebald and Véron, 1963).

But it was not until 1977 that the importance of *Campylobacter* in human disease came to wide attention with the publication of Skirrow (1977) in which he reported a 7% incidence of *Campylobacter* positive stool samples in diarrhoeal patients using a selective isolation protocol refined from earlier work (Butzler *et al.*, 1973). In the years that have followed, *Campylobacter* and particularly the most common human disease causing strain, *C. jejuni*, has been the subject of intense research interest and is currently the target of multiple national and international research initiatives.

1.3.2 *C. jejuni* exerts a large burden of disease

The significance of *C. jejuni*, and other *Campylobacter* spp., in human disease is now widely recognised. In fact, *Campylobacter* is the leading cause of food poisoning in the UK and the developed world as a whole. It accounts for around three quarters of all laboratory confirmed cases of foodborne disease in the UK which leads to an economic cost estimated by the Food Standards Agency (FSA) as £900 million and causes around 100 deaths each year (Wearne, 2013).

C. jejuni, and other *Campylobacter* species, are extremely common in farmed animals with poultry, and chickens in particular, being the most common source of transmission to humans. In the UK some 70-80% of intensively reared chicken meat and almost 100% of free range or organically reared chicken meat carries *C. jejuni* (Colles *et al.*, 2008, EFSA, 2010), although freezing substantially limits bacterial viability meaning frozen chicken products are substantially safer (Jacobs-Reitsma *et al.*, 2008). Recently the UK has been pursuing a name and shame approach to *C. jejuni* in supermarket chicken and early indications are that this is driving a modest reduction in levels of *C. jejuni* in the foodchain (FSA, 2016).

The high incidence of *Campylobacter* in the food chain is doubtless partly responsible for its importance in human disease, accounting for around three quarters

of laboratory confirmed intestinal infection and a third of the costs arising from foodborne illness in the UK (EFSA, 2010). *Campylobacter* infection manifests as a gasteroenteritis with symptoms similar to those caused by other bacteria such as *Salmonella* and is known as campylobacteriosis. Because of the similarities between campylobacteriosis and other conditions this disease can only be definitively diagnosed by stool sample analysis (Fitzgerald *et al.*, 2008).

Campylobacteriosis is rarely life threatening, however in a small number of cases patients go on to develop Guillain Barré syndrome (GBS) two to three weeks after the initial infection. Named after its discoverers, GBS is characterised by an acute flaccid paralysis (Yuki and Hartung, 2012). The symptoms progress over a period of hours, starting with numbness of the extremities, spreading to loss of muscle control in the limbs and ending with a total loss of muscle tension throughout the body. Without access to advanced medical facilities this paralysis is usually fatal as the patient becomes unable to breath. Rapid transfer of sufferers to a hospital, however, results in survival of around 95% of patients although a significant number never make a complete recovery with permanent disability occurring in around 15% of cases (Hughes *et al.*, 2007).

In contrast to its pathogenic effect in humans, *C. jejuni* has frequently been described as a harmless commensal of chicken (e.g. Fouts *et al.*, 2005, Coward *et al.*, 2008, Guerry and Szymanski, 2008, Gutiérrez-Martín *et al.*, 2011, Hermans *et al.*, 2011). However, more recent evidence has found evidence of inflammation, damage to the gut mucosa, and diarrhoea in chickens resulting from *C. jejuni* colonisation (Humphrey *et al.*, 2014) although this effect seems to be breed dependent.

1.4 Physiology and taxonomy of *C. jejuni*

C. jejuni is a spiral shaped, Gram negative member of the epsilonproteobacteria. *C. jejuni* cells are typically between 0.5 and 5.0 µm long and flagellated at one or both poles (Snelling *et al.*, 2005). Although able to tolerate atmospheric oxygen levels for short periods of time, it is primarily microaerophilic, and grows best at temperatures between 37°C and 42°C (Hazeleger *et al.*, 1998). In the laboratory *C. jejuni* is usually cultured on either blood agar or Mueller-Hinton agar (MHA) plates, but must be maintained under microaerobic conditions either in a gas jar, in a culture cupboard with controlled O_2 levels, or in a microaerobic workstation (Hazeleger *et al.*, 1998, Macé *et al.*, 2015).

1.4.1 Other species in genus *Campylobacter*

Many species of *Campylobacter* have been associated with a range of diseases, both in humans and domesticated animals. The most commonly encountered species are *C. jejuni* and *C. coli*, which together are the leading cause of food-borne illness in the developed world, primarily due to their ability to colonise the caecum of chickens and then enter the food chain via raw chicken meat, although both species are found in a range of domestic and wild animals. Evidence is growing that *C. ureolyticus* also makes up a significant proportion of *Campylobacter* related gastroenteritis, with one study suggesting it is present in 1 in 4 cases of *Campylobacter* related diarrhoea (O'Leary *et al.*, 2009). This prevalence is believed to stem from the presence of *C. ureolyticus* in cattle (O'Donovan *et al.*, 2014). *C. ureolyticus* is also associated with other diseases including Crohn's disease, irritable bowel syndrome (IBS), and various forms of lesion and ulcer (Burgos-Portugal *et al.*, 2012). *C. hyointestinalis* was first identified in pigs with proliferative enteritis but also isolated from cattle faeces and the intestines of hamsters (Gebhart *et al.*, 1985).

Several *Campylobacter* species are found in wild birds: *C. lari* is primarily found in seagulls but is also disease-causing and occasionally fatal in humans (Benjamin *et al.*, 1983, Martinot *et al.*, 2001, Werno *et al.*, 2002), *C. volucris* was isolated from black-headed gulls in Sweden (Debruyne, Broman, Bergstro, On and Vandamme, 2010), and *C. subantarcticus* was isolated from wild birds in the sub-Antarctic region (Debruyne, Broman, Bergström, Olsen, On and Vandamme, 2010). But *Campylobacter* species are not limited to birds; *C. insulaenigrae* was initially isolated from common seal (*Phoca vitulina*) and harbour porpoise (*Phocoena phocoena*) (Foster *et al.*, 2004) but has since been found in northern elephant seals (*Mirounga angustirostris*) (Stoddard *et al.*, 2007) and South American sea lions (*Otaria flavescens*) (González *et al.*, 2011), suggesting that it inhabits a broad range of marine mammal hosts. *C. iguaniorum* was discovered in rectal or cloacal swabs from a range of reptiles (Gilbert *et al.*, 2015) and *C. peloridis* (Debruyne *et al.*, 2009) was identified in shellfish. These species are considered to be part of the *C. lari* subgroup (Miller *et al.*, 2014).

There are also other diseases associated with *Campylobacter*. *C. concisus*, *C. curvus* and *C. gracilis* have all been isolated from gingivitis patients and occasionally from wound sites (Tanner *et al.*, 1981; 1984, Merriam *et al.*, 2003), although *C. concisus* and *C. curvus* have also been found associated with intestinal disease (Abbott *et al.*, 2005, Kaakoush and Mitchell, 2012). *C. fetus* is associated with spontaneous abortion in sheep and cattle (see section 1.3.1). *C. hominis* is unusual among *Campylobacter*s because it is frequently isolated from faecal samples of both normal and diarrhoeic humans suggesting that it can exist as a commensal in humans (Lawson *et al.*, 1998).

1.4.2 The *C. jejuni* glycome

The surface of *C. jejuni* is coated in a range of glycolipids and glycoproteins. These come primarily in four classes: the lipooligosaccharide, the capsular polysac-

charide, the glycosylated flagellin proteins, and the *N*-linked glycoproteins (see figure 1.2). All but the last of these are highly variable between species and thought to play an important role in immune recognition, bacteriophage attachment and other surface functions (Karlyshev *et al.*, 2005). This variability exists both as between strain variation in the genes responsible and within strain variation in phase variable loci.

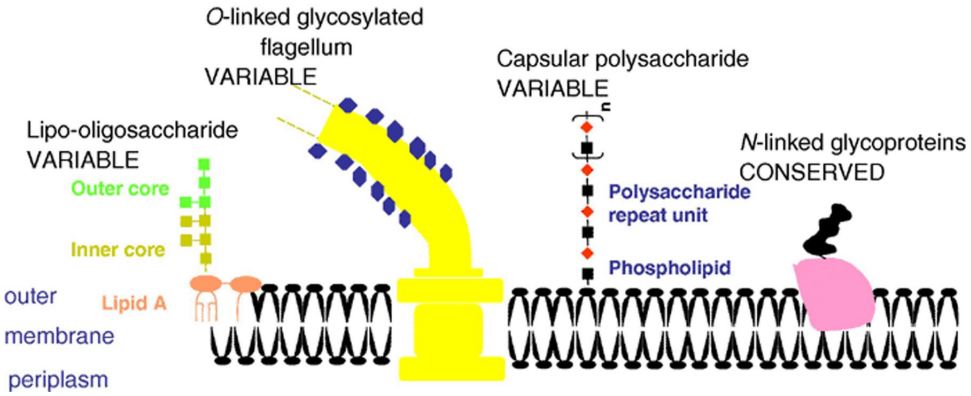

Figure 1.2: The *C. jejuni* glycome
The major classes of surface glycolipid and glycoprotein in *C. jejuni*. 'Variable' here refers to molecules that are highly variable between strains rather than necessarily phase variable, however there are phase variable genes modifying all three of the variable classes in strain NCTC 11168. Figure is reproduced from Karlyshev *et al.* (2005) with permission.

Lipopolysaccharides (LPSs) are long chain sugars linked to a lipid root that is anchored in the outer membrane of Gram-negative bacteria. Typically these molecules are highly variable and have several distinct regions; there is a lipid root (e.g. lipid A), then a largely unchanging saccharide linker region, followed by several saccharide repeats and a final long carbohydrate chain. This final chain is absent in *C. jejuni* and the reduced LPS is normally referred to as lipooligosaccharide (LOS) to distinguish it from the more normal bacterial structure. The genes involved in LOS synthesis are physically located close to each other in the genome in a cluster of genes starting at *cj1133* (*waaC*) in NCTC 11168 but are highly variable between strains (Karlyshev *et al.*, 2005). These genes construct the LOS by attaching one moiety at a time, starting from lipid A. In strain NCTC 11168, the terminal moiety is a β-1,3 Galactose attached by the phase variable gene *cj1139* (*wlaN*), this alters the final structure from mimicking ganglioside GM_1 to ganglioside GM_1a (Linton *et al.*,

2000). Phase variable changes in ganglioside mimicry also occur in other strains. In *C. jejuni* strain 81-176 changes in *cgtA* expression produces a change from ganglioside GM_2 mimicry the ganglioside GM_3 mimicry (Guerry and Szymanski, 2002). Because GBS is thought to result from auto-immune responses involving these gangliosides, these changes may be important to disease progression (Nachamkin *et al*., 1998; 2002, Yu *et al*., 2006).

Until the year 2000, *C. jejuni* was generally considered to be uncapsulated (Moran and Penner, 1999, Karlyshev *et al*., 2008) although Chart *et al*. (1996) had isolated polysaccharides from *C. jejuni* that they identified as likely of capsular origin. This changed when the publication of the first genome sequence (Parkhill *et al*., 2000) showed that the *C. jejuni* genome contained genes homologous to the *kps* gene cluster involved in capsular polysaccharide synthesis in *Escherichia coli* and other bacteria (Pavelka *et al*., 1991, Whitfield, 2006). Based on this genetic evidence knock-out strains of the *kpsM*, *kpsC*, and *kpsS* genes were analysed and all were found to result in the loss of serotype (Karlyshev *et al*., 2000). This demonstrated that the capsule, and not the LOS – as previously thought – was being recognised by the serotyping methods (Karlyshev *et al*., 2000). Accordingly, the capsular polysaccharides (CPS) is the most immunogenic antigen on the surface of *C. jejuni*. Shortly afterwards, the use of Alcian blue dye enabled the detection of the capsule by electron microscopy (Karlyshev *et al*., 2001).

As with the LOS, the CPS biosynthesis genes are primarily closely located within a particular region of the genome. There are 38 genes in this biosynthesis locus (Karlyshev *et al*., 2005) located between *cj1414* and *cj1443*. The structure of the CPS is highly variable between *C. jejuni* strains and frequently contains several phase variable loci that produce additional population level diversity. The CPS is further modified by two MeO*P*N transferases discussed in section 1.7.1 below. The *C. jejuni* flagellum is primarily composed of the flagellin FlaA, with a smaller proportion of highly similar FlaB, and a number of other proteins that make up the

hook and cap. FlaA is highly glycosylated with a range of moeties that are again variable between *C. jejuni* strains (Ulasi *et al.*, 2015). Differences in glycosylation of the flagellum are known to influence motility and auto-agglutination (Karlyshev *et al.*, 2002, Guerry *et al.*, 2006, van Alphen *et al.*, 2008) and may influence phage binding (Baldvinsson, 2014).

1.4.3 The NCTC 11168 isolate of *C. jejuni*

C. jejuni strain NCTC 11168 is one of the most commonly used laboratory strains of *C. jejuni* and the primary strain used in the following work. NCTC 11168 was isolated from stool samples taken from a human subject suffering from gastric enteritis in 1977 by Martin Skirrow (Skirrow, 1977) and carried out at the Royal Worcester Infirmary in the UK. As one of the most widely used laboratory strains it was a natural candidate for genome sequencing and became the first *Campylobacter* strain to have its genome sequenced (Parkhill *et al.*, 2000). Seven years after the original sequencing project, NCTC 11168 was re-analysed and re-annotated (Gundogdu *et al.*, 2007). Since its isolation the strain used to generate the genome sequence, and likely most lab isolates, have diverged from the original isolate with differences seen in motility, morphology, and in expression of genes related to metabolism and respiration (Gaynor *et al.*, 2004).

The genome of NCTC 11168 consists of roughly 1.65 million base pairs, encoding for 1643 open reading frame (ORF)s, with no plasmids. This is similar in length to other strains that have since been sequenced (Pearson *et al.*, 2007, Zhang *et al.*, 2010, Shyaka *et al.*, 2015). *C. jejuni* has an AT rich genome (GC content is around 25-30%, depending on strain), but despite this AT-richness, *C. jejuni* genomes contain a large number of poly-G tracts. In NCTC 11168 there are 29 poly-G tracts consisting of between 7 or more guanine residues in a short tandem repeat sequence. It is

these poly-G tracts that form the molecular base for phase variation in *C. jejuni* (see section 1.5).

1.4.4 Genetic features of *C. jejuni*

This first genome sequence revealed a substantial amount of new information about the genome *C. jejuni*. *C. jejuni* appears to use the normal bacterial codon translation table but, unlike other bacteria, the genomes of *C. jejuni* are not primarily organised into operons (Parkhill *et al.*, 2000). *C. jejuni* contains just three sigma factors σ^{28} (*fliA*), σ^{54} (*rpoN*), and σ^{70} (*rpoD*) (Parkhill *et al.*, 2000, Mittenhuber, 2002). σ^{70} is responsible for regulation of a range of housekeeping genes (Wösten, 1998, Petersen *et al.*, 2003) but also plays a role in the regulation of the other two σ-factors (Carrillo *et al.*, 2004) which are involved in the regulation of flagellar and motility related genes (Wösten, 1998, Carrillo *et al.*, 2004).

In common with many bacteria, *C. jejuni* growth is limited by iron availability and so regulating iron uptake is an important part of its pathogenicity and growth (van Vliet and Ketley, 2001). The Fur family of metalloregulators are an important group of iron sensing proteins that regulate a range of iron uptake pathways. They are found in a wide range of bacteria (Lee and Helmann, 2007) including *C. jejuni* (Wooldridge *et al.*, 1994, van Vliet *et al.*, 2002). Although *fur* is essential in some bacteria, *fur* mutants are viable in *C. jejuni* although they show reduced growth compared to wild type in both iron replete and iron restricted conditions (van Vliet *et al.*, 1998). Although Fur is the principal iron regulatory protein in *C. jejuni*, it is joined by another member of the Fur family, PerR, which is homologous to an iron regulatory system found in *Bacillus subtilis* (Parkhill *et al.*, 2000).

Parkhill *et al.* (2000) identified twelve two component regulatory systems in the original NCTC 11168 genome sequence, although these systems were already known

to exist in *C. jejuni* (Brás *et al.*, 1999). Two component systems are widespread in prokaryotes and operate by a system of interaction between two parts (Beier and Gross, 2006, Mitrophanov and Groisman, 2008). The first is often embedded in the membrane and typically operates as a histidine kinase, the second is a response regulator that when phosphorylated by the first protein then produce a response within the cell either by activating downstream processes or directly binding to the DNA and acting as transcription factors. In the classic example of a two-component system the first protein triggers the second in response to an extra-cellular signal and frequently also acts as the phosphotase responsible for de-phosphorylating the response regulator when the signal has ended (Mitrophanov and Groisman, 2008). In *C. jejuni*, two component systems are involved in a range of functions including temperature response, motility, flagellum expression, and biofilm formation (Brás *et al.*, 1999, Dasti *et al.*, 2010, Svensson *et al.*, 2015). Thus, in *C. jejuni*, phase variation sits alongside a range of more conventional regulatory mechanisms that allow the bacterium to respond to its environment.

1.4.5 Sequence typing and ST-complexes

Because morphological differences between bacterial strains provide only a crude representation of the genetic variability and evolutionary relationship between bacteria, there has been a need for more precise methods to distinguish between strains. While typing methods such as serotyping (Moran and Penner, 1999) and bacteriophage typing (Grajewski *et al.*, 1985) provided improved separation of strains, it was not until the advent of multi-locus sequence typing (MLST) that more accurate evolutionary relationships could be established. MLST relies on the amplification of a small number of housekeeping genes common to the species of interest. These genes are sequenced and numbered according to identified alleles. The combination of alleles is then assigned a sequence type (Maiden *et al.*, 1998).

This technique has been successfully applied to *C. jejuni* with the seven genes *aspA*, *glnA*, *gltA*, *glyA*, *pgm*, *tkt*, and *uncA* (Dingle *et al.*, 2001, Colles and Maiden, 2012).

The sequence type usefully identifies lineages of bacteria, however it is frequently the case that there are many sequence types (STs) that are separated by a variations in just one or two alleles (see figure 1.3). These clusters are referred to as clonal complexes, or ST-complexes. Despite originally being informally defined they have proved a useful level of analysis (Dingle *et al.*, 2002, Colles *et al.*, 2003) and have now been placed on a more formal footing (Didelot and Falush, 2007). Although whole genome sequencing offers a new level of detail, MLST remains an important and useful means of classifying bacteria especially between the meroclone and species level and whole genome sequencing confirms the evolutionary relationships thought to underpin MLST (Maiden *et al.*, 2013, Pérez-Losada *et al.*, 2013)

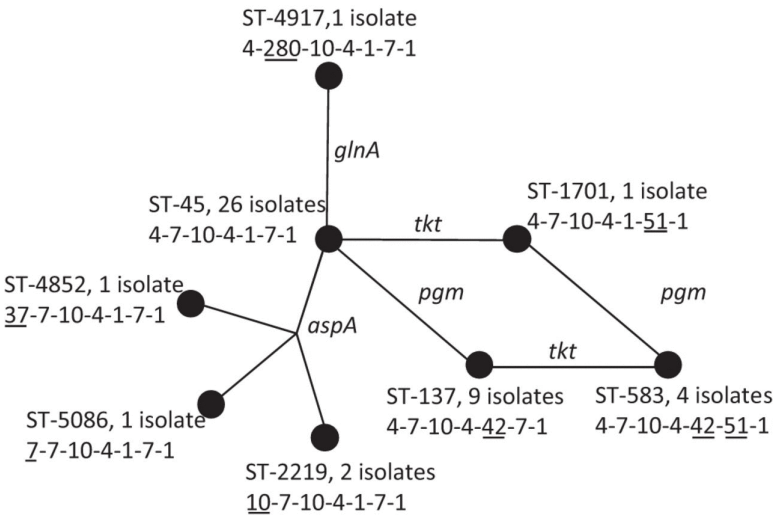

Figure 1.3: Illustration of ST complex
Example of an ST-complex in *C. jejuni*. Data is taken from Oxfordshire human disease isolates collected in 2008. These isolates cluster around the central ST-45 type to which the majority of isolates belong with the other ST types differing by one or two alleles from this central type. Allele types are given in the order *aspA-glnA-gltA-glyA-pgm-tkt-uncA* and allele types are underlined if they differ from the ST-45 types. A notable point about this tree is that it does not follow the progressively dividing tree structure expected from vertical evolution. Image taken from (Colles and Maiden, 2012), and used under the Creative Commons Attribution License.

1.5 Phase variation in *C. jejuni*

The genome of *C. jejuni* contains a significant number of poly-G tracts consisting of 7 or more tandemly-arranged G-residues. These tracts are highly mutable and occur more often than expected by chance (Lin and Kussell, 2012). Loss or gain of a single G from these tracts occurs at a rate of once every hundred to a thousand divisions – a mutation rate three to five orders of magnitude higher than that of the genome in general. Levinson and Gutman (1987) proposed that the mechanism behind this mutability was slipped-strand mispairing, and the available evidence supports this view (Bayliss, 2009). In many cases, these poly-G tracts are located within the coding region of proteins and thus the gain or loss of a single base pair can induce frameshift mutations, typically truncating the protein and producing a non-functional product. However, the poly-G tract remains intact and continues to mutate with high frequency so that the tract can undergo a reverse mutation with high probability and restore the functional protein. These in-frame poly-G tracts represent a mechanism for rapid, reversible, inheritable switching between two gene expression states (see section 1.2 above).

1.5.1 Phase variation in other *Campylobacter* species

Variable poly-G tracts, similar to those in *C. jejuni*, have been identified in other *Campylobacter* species (e.g. Pearson, 2013, Kienesberger *et al.*, 2014). Slipped-strand mispairing, however, is not the only mechanism for phase variation in *Campylobacter*: *C. fetus* uses a system of nested DNA inversions to produce a large repertoire of different surface layer proteins (reviewed in Dworkin and Blaser, 1997). This mechanism is *recA*-dependent and works by a series of consecutive inversion events occurring within the reading frame of the protein. These inversions operate on regions defined by pairs of inverted repeats.

1.5.2 Phase variable loci of NCTC 11168

The NCTC 11168 strain of *C. jejuni* is commonly used for research purposes and was the first *C. jejuni* strain to have a whole genome sequence produced (Parkhill *et al.*, 2000, Gundogdu *et al.*, 2007). This sequencing allowed the identification of 29 potentially variable poly-G tracts, and confirmed variability in 21 of these (as well as in a T4/5 tract and a C1/2 tract). 27 out of these 29 SSRs are located within or close to coding regions, with three associated with pseudogenes. The majority of the remainder are associated with genes involved with capsular, LOS or flagella modification. Eight phase variable genes have been putatively identified as linked to flagellar glycosylation, three are of the *cj1318*-, or *maf*-, family which has been shown to modify glycosylation in the 81-176 strain (Guerry and Szymanski, 2002) while the remainder are of the 617 family. Table 1.1 lists all 29 tracts and information about the genes or ORFs that they are located in or nearest to. For convenience, these loci will be referred to throughout this thesis by the names given in this table which are derived from the associated ORF.

cj0031 encodes a type IIG restriction-modification (R/M) enzyme which methylates the adenine in sequences matching 5′-CCYGA-3′in the ON state (Anjum *et al.*, 2016). As with other type IIG R/M systems both the restriction and modification enzymes are part of the same sequence chain and so both functions are switched together. Phase variation in R/M systems may help prevent build up of resistance in bacteriophages and thus increase population level resistance to infection producing frequency-dependent selection for the OFF state (Bayliss *et al.*, 2006). However, there is also evidence of selective changes in *cj0031* expression during passage through chickens (Bayliss *et al.*, 2012) which is unlikely to result from this effect. The concept of a 'phasevarion' of genes whose expression is influenced by the methylation effect of R/M systems has gained considerable interest in recent years (Srikhanta *et al.*, 2010). Anjum *et al.* (2016) used RNASeq to identify multiple genes

PV Locus	Location	ON tract	Function	References
cj0031	48995	G9	Type IIG R/M system	Anjum *et al.*, 2016
cj0045c	65747	G11	Iron uptake system[1]	Parkhill *et al.*, 2000, Gundogdu *et al.*, 2007
cj0046	67707	(G11)	Pseudogene (sodium sulfate transmembrane transport)	Parkhill *et al.*, 2000, Gundogdu *et al.*, 2007
cj0170	167292	G8	Influences motility	Artymovich *et al.*, 2013
cj0275 (ClpX)	252477	G8	Protein degradation pathway	Cohn *et al.*, 2007
cj0565 (upstream)	527377	(G10)	Pseudogene	Parkhill *et al.*, 2000, Gundogdu *et al.*, 2007
cj0617	577585	G10	Flagellar[2] modifying transferase[1] (617 family)	Parkhill *et al.*, 2000, Gundogdu *et al.*, 2007
cj0628 (CapA)	588368	G9	Cell adhesion (Lipoprotein autotransporter)	Ashgar *et al.*, 2007
cj0676	628169	G9	Pseudogene (potassium transport ATPase)	Parkhill *et al.*, 2000, Gundogdu *et al.*, 2007
cj0685c (CipA)	639006	C9	Cell invasion protein	Lynett, 1999, Javed *et al.*, 2010
cj0742 (downstream)	695942	N/A[3]	Psuedogene (outer membrane protein)	Parkhill *et al.*, 2000, Gundogdu *et al.*, 2007
cj1139c	1074192	G8	LOS-modifying β-1,3 galactosyltransferase	Linton *et al.*, 2000
cj1144c	1079739	G10	LOS-modifying[2]	Parkhill *et al.*, 2000, Gundogdu *et al.*, 2007
cj1295	1227120	G9	Flagellar modifying glycosyltransferase	Hitchen *et al.*, 2010
cj1296	1228590	G10	Flagellar modifying[2]	Parkhill *et al.*, 2000, Gundogdu *et al.*, 2007
cj1305c	1234921	G9	Flagellar modifying[2] (617 family)	Parkhill *et al.*, 2000, Gundogdu *et al.*, 2007
cj1306c	1236160	G9	Flagellar modifying[2] (617 family)	Parkhill *et al.*, 2000, Gundogdu *et al.*, 2007
cj1310c	1240299	G9	Flagellar modifying[2] (617 family)	Parkhill *et al.*, 2000, Gundogdu *et al.*, 2007
cj1318 (maf1)	1246845	G8	Motility accessory factor, flagellar glycosylation	Karlyshev *et al.*, 2002, van Alphen *et al.*, 2008
cj1321 (upstream)	1250805	(G10)	Flagellar modifying[2]	Parkhill *et al.*, 2000, Gundogdu *et al.*, 2007
cj1325	1253668		Flagellar modifying[2] methyltransferase[1]	Parkhill *et al.*, 2000, Gundogdu *et al.*, 2007
cj1335 (maf4)	1263674	G8	Motility accessory factor, flagellar glycosylation	Karlyshev *et al.*, 2002, van Alphen *et al.*, 2008
cj1342 (maf7)	1275366	G9	Motility accessory factor	Karlyshev *et al.*, 2002
cj1420c	1353746	G9	Capsular[2] methyltransferase[1]	Parkhill *et al.*, 2000, Gundogdu *et al.*, 2007
cj1421c	1355957	G9	CPS MeO*P*N transferase	McNally *et al.*, 2007, Sørensen *et al.*, 2011; 2012
cj1422c	1357889	G9	CPS MeO*P*N transferase	McNally *et al.*, 2007, Sørensen *et al.*, 2011; 2012
cj1426c	1357889	G10	CPS methyltransferase	Sørensen *et al.*, 2012, Sternberg *et al.*, 2013
cj1429c	1363821	G10	Capsular modification[2]	Parkhill *et al.*, 2000, Gundogdu *et al.*, 2007
cj1437c	1374135	G9	Capsular modification[2]	Parkhill *et al.*, 2000, Gundogdu *et al.*, 2007

[1] Putative function deduced by sequence similarity
[2] Putative function deduced from genomic position
[3] Not included in 28-locus fragment analysis assay (see section 2.3)

Table 1.1: The phase variable loci of strain NCTC 11168
All 29 Poly-G tracts of length 7 or more are shown. Tracts are named for the ORF they are located within, or the closest ORF if they are not located within a reading frame. Names given in parentheses are alternative names for the same locus. ON lengths are putative for most tracts based on the length that gives the longest reading frame. ON lengths shown in parentheses are for intergenic tracts and tracts located in pseudogenes. Intergenic tracts have been assigned an arbitrary ON length for use in some analyses. For pseudogenes the length that produces the longest run before a stop codon is chosen as the 'ON' state.

which showed differential expression dependent on the Δ*cj0031* strain but found little association between these differences and the presence of *cj0031* methylation sites. The differences in expression appeared to be responsible for phenotypic differences in adhesion to, and invasion of, Caco-2 cells and differences in biofilm formation between the *cj0031*-OFF and *cj0031*-ON and comparable differences in *cj0031* knockout, and complemented knockout, strains (Anjum *et al.*, 2016).

cj0045c is homologous to iron-transport proteins, however the location of the tract is not close to the start of the open reading frame, where it would be expected to create phase variable expression but rather located at the end of the reading frame. Changes in repeat number result in termination before or after the start codon of the downstream gene *cj0044c*. It has been suggested, therefore, that rather than influencing the expression of *cj0045c*, the tract may influence expression levels of *cj0044c* by inhibiting the correct binding of the transcription complex to the *cj0044c* initiation codon. *cj0046* is a pseudogene containing multiple frameshift mutations. The functional homologues of this gene are involved in sodium sulfate transmembrane transport and this site has been used as a target for insertion of complementation cassettes (Thomas *et al.*, 2011, Anjum *et al.*, 2016).

As with *cj0045c*, the tract in *cj0170* is located close to the end of the reading frame resulting in a maximum truncation of 22 amino acids. Surprisingly, this maximally truncated version has been found to be the functional length of the protein (Artymovich *et al.*, 2013) with a marked effect on the motility of the bacterium. The maximally truncated length is also associated with colonisation and disease in a mouse model (Kim *et al.*, 2012) but the mechanism of action has not been established.

cj0275 is a phase variable homologue of the ClpX protein which is an AAA+ protease that forms part of the ClpPX pathway for protein degradation (Baker and Sauer, 2012) and may also have independent chaperone activity (Wawrzynow *et al.*, 1995).

It is unclear what functional benefit variation in this gene might have and there is evidence that the ClpP and ClpX proteins are required for growth at 42°C (Cohn *et al.*, 2007). Supporting this, Lango-Scholey *et al.* (2016) did not detect any *cj0275*-OFF colonies among 390 colonies recovered from passage of *C. jejuni* in chicken, whereas all 27 other PV loci showed some variation in ON/OFF state during the course of the experiment. This study indicates that there is strong selective pressure for the ON state of this gene during *in vivo* growth in chicken.

The '*cj0565*' tract is, in fact, located upstream of *cj0564* and downstream of the pseudogene *cj0565*. No functional association of this tract is known. *cj0617* is the eponymous member of the 617-family of genes that are thought to be involved in flagellar glycosylation. This attribution is based on the presence of most genes of this family within a flagellar region of the genome but *cj0617* itself is not located within this region (Parkhill *et al.*, 2000).

Also known as CapA, *cj0628* is one of two of the poly-G tracts in NCTC 11168 which have an associated poly-A/T tract. In this case, the poly-G tract is proceeded by a poly-T tract of length 5 which may also exhibit some variability adding to the overall variability of this gene (Parkhill *et al.*, 2000). CapA was identified as an autotransporter lipoprotein by bioinformatic analysis and then demonstrated to influence Caco-2 epithelial cell invasion *in vitro* and adhesion to chicken intestinal cells *in vivo* (Ashgar *et al.*, 2007). CapA is one of two closely homologous Cap proteins with identical C-terminus regions. The other gene *capB* is not expressed in NCTC 11168 because of a frameshift mutation associated with a 7T homonucleotide tract; however, a single deletion, or a double insertion, to this T-tract would produce a functional version of CapB (Ashgar *et al.*, 2007). Such poly-T tracts have been less closely studied than poly-G tracts but appear to be more stable (Bayliss, unpublished data) and so any variability produced by this tract is likely to be lower than that in "true" phase variable genes.

cj0676 is another pseudogene, with similarities to potassium-transporting ATPases. However, there is evidence that this gene is functional in some virulent isolates of NCTC 11168 suggesting that this is a lab-acquired mutation (Cooper *et al.*, 2013). Named as *Campylobacter* invasion protein A (CipA), *cj0685c* has been identified as associated with cell invasion in Caco-1, Caco-2 (Lynett, 1999), and INT-407 cells (Javed *et al.*, 2010). Bioinformatic comparison to proteins of known function identifies it as a putative sugar transferase. This gene is also notable as the only one in which the tract is poly-C rather than poly-G in the direction of translation.

The next tract is located 20bp downstream of *cj0742*, which is a pseudogene with homology to known outer membrane proteins, and 470bp downstream of a 16S ribosomal rRNA. It is not clear how this tract could have any functional role. *cj1139c* (wlaN) is a β-1,3 galactosyltransferase which modifies the LOS to resemble the GM_1 ganglioside of humans (Linton *et al.*, 2000). Molecular mimicry of gangliosides is thought to underlie Guillain-Barré syndrome so the phase variable state of this gene is potentially involved in the progression of disease from unpleasant but relatively harmless gastroenteritis to life-threatening GBS (Yu *et al.*, 2006).

The exact function of *cj1144c* is unknown, but it has been suggested to be involved in LOS modification based on its genomic location. There is weak evidence that it is involved in phage sensitivity to F341 (Baldvinsson, 2014), which is a flagellar binding phage so this assumption of function based on location may be flawed. *cj1144* is the second poly-G tract associated with a poly-A/T tract, this time a 9A tract in the sequenced genome. *cj1295* has been more definitively identified as a flagellar glycosyltransferase. It attaches a di-O-methylglyceroyl-modified version of pseudaminic acid to the major flagellin protein, FlaA (Hitchen *et al.*, 2010). Serial passage through a mouse model resulted in an enrichment of the 9G ON variant of *cj1295* (Jerome *et al.*, 2011). The same mouse passage experiment also found enrichment in the adjacent PV locus *cj1296* also thought to be a flagellar modifying gene.

The next three PV loci, *cj1305c*, *cj1306c* and *cj1310c* are also located within this flagellar locus and are thought to be glycosyltransferases. They have fairly high homology (> 60% amino acid sequence similarity) to each other and slightly weaker similarity to *cj0617* and are considered to be part of the same 617 family of genes. The exact function of these genes is not known. *cj1306c* showed statistically significant enrichment of the ON state after serial passage through a mouse model (Jerome *et al.*, 2011).

There are three PV *maf* (motility-accessory factor) in NCTC 11168. *cj1318* (*maf1*) and *cj1335* (*maf4*) have 98% amino acid sequence similarity to each other while *cj1342c* (*maf7*) is more distantly related. The alternative names for these genes come from experimental work showing that the homologous genes in *C. jejuni* strain 108 influence the motility of the bacterium (van Alphen *et al.*, 2008). In strain 81-176, the homologous genes influence the glycoforms of sugars attached to the flagella (Guerry *et al.*, 2006) and it may be that this modification provides the basis for the change in motility. As well as the influence on motility, these genes also impact autoagglutination (Guerry *et al.*, 2006).

Between the *maf1* and *maf4* genes are two more PV loci, *cj1321* and *cj1325*, which are also considered likely to modify flagellar proteins. The tract for *cj1321* is not located within the reading frame but 46bp upstream. The positioning of this tract means that it may impact the binding of transcription factors and lead to differences in transcriptional expression. *cj1321* is homologous to amino-transferases in other species, while *cj1325* is a putative methyltransferase with just over 50% protein sequence similarity to *cj0170* and thus may modify motility in the same way as *cj0170* (see above). The cluster of five genes *cj1321-cj1325* is strongly associated with livestock strains of *C. jejuni* (Champion, 2005), however experimental investigation of chicken colonization with knockout strains found that *cj1324* appears to be the important factor in this association (Howard *et al.*, 2009).

cj1420c is a methyl-transferase positioned between the phosphoramidate biosynthesis pathway and the two phase variable phosphoramidate transferases *cj1421c* and *cj1422c*. *cj1421c* and *cj1422c* are of particular relevance to this project and so are discussed in detail in section 1.7.1 below. The final gene identified as undergoing shifts during serial passage through a mouse model by Jerome *et al.* (2011) is *cj1429c*. The gene is part of a capsular locus but has no known function. Also part of this capsular locus is the final PV locus, *cj1437c*, which has sequence similarities to aminotransferases based on bioinformatic analysis.

1.6 Bacteriophages of *C. jejuni*

The term bacteriophage (commonly shortened to 'phage') was first used in 1917 by D'Hérelle and refers to viruses that infect bacteria, although the British scientist Frederick Twort discovered them independently two years earlier (Duckworth, 1976). Believed to be the most numerous form of organic replicator in existence, bacteriophages reproduce by entering the bacterial cell and hijacking the replicative machinery of the cell in order to create further copies of themselves. Replication can either occur immediately or the phage can integrate into the DNA of the host and lie dormant, potentially being multiplied by the usual host replicative processes, until later becoming active (this is referred to as lysogeny). When active, bacteriophages typically produce hundreds of copies of themselves before the cell bursts releasing the newly synthesized phages, in what is referred to as a lytic burst. Phages that can reproduce by both mechanisms are referred to as temperate, whereas phages that cannot undergo lysogeny are referred to as lytic.

The first discovered *Campylobacter* bacteriophages were identified in *C. coli* and *C. fetus* (Connerton *et al.*, 2011), and by 1985 sufficient *Campylobacter* bacteriophages were known to be used as a typing system for *C. coli* and *C. jejuni* (Grajewski *et al.*,

1985). Some phages were found because they cause autoagglutination and interfere with serotyping methods (Ritchie *et al.*, 2016).

Since then many more distinct *C. jejuni* bacteriophages have been identified (Connerton *et al.*, 2004, El-Shibiny *et al.*, 2005, Hansen *et al.*, 2007). Recently, simple screening experiments aimed at isolating bacteriophages from farmyard samples have identified large numbers of novel bacteriophage (e.g. Sørensen *et al.*, 2015) indicating the existence of large numbers of uncharacterised *C. jejuni* bacteriophage. These phages vary in their host ranges with many *C. jejuni* strains able to resist infection by one or more bacteriophage strain (Sørensen *et al.*, 2015) but most are drawn from just one of the 18 widely recognised phage families, the *Myoviridae* (Connerton *et al.*, 2011). The *Myoviridae* have characteristic icosahedral heads connected to a tail of varying length and then a series of tail fibres involved in binding to host surface structures (termed receptors). The genetic material of *Myoviridae* phages is maintained as linear double stranded DNA (Ackermann, 2003).

C. jejuni bacteriophages studied so far fall into two broad classes: those that interact with the flagellum and those that attack via the capsule. Flagellar phages can be shown to require a functional flagellum, which allows bacterial motility, in order to attack the bacterium (Baldvinsson, 2014) and electron microscopy provides clear images showing the tail fibres binding to the flagellum but the exact mechanism of attachment and invasion has not yet been elucidated.

1.7 Selection of target gene and selective agents for a cyclical assay

One of the primary focuses of this work is the development and modelling of a cyclical selection assay in which alternating selective pressures can be applied to a

single phase variable gene so the ON and OFF states are sequentially selected for (see figure 1.4). In order to create this assay the first step is to identify a target phase variable gene together with two forms of selection that have opposite effects on that gene. This section discusses the chosen gene – *cj1421c* – and the possible methods of selection.

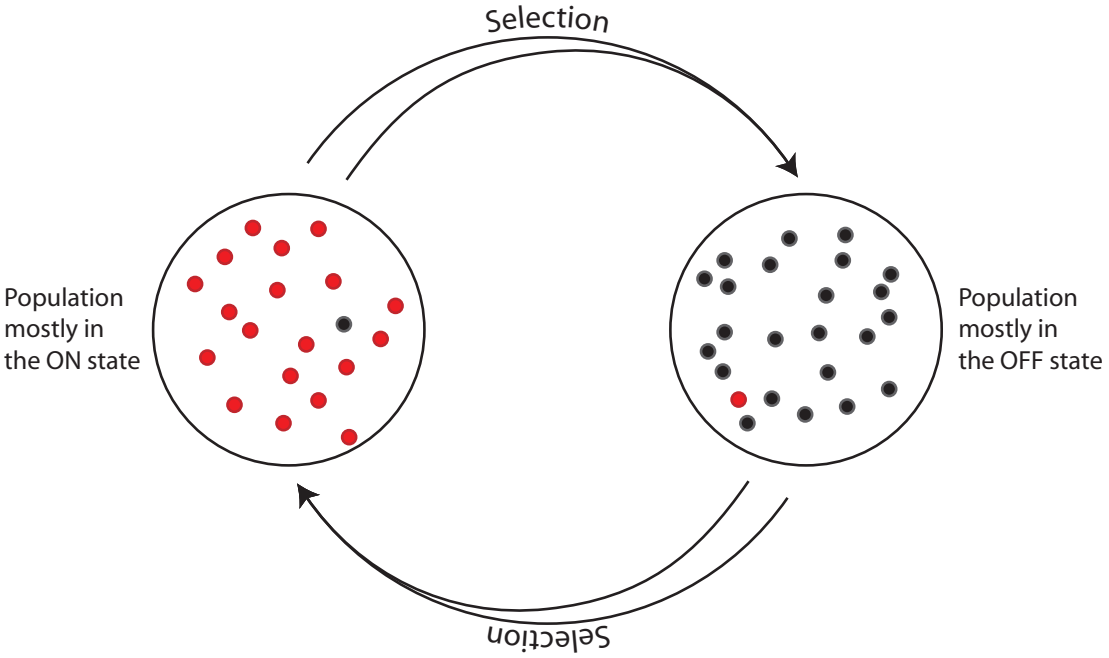

Figure 1.4: The concept behind the cyclical selection assay
The concept of a cycle selection assay is to create a cycle in which selection forces the population to fluctuate between the ON and OFF states of a gene. To do this, requires a means to take a population with a phase variable gene mostly in the ON state, apply some kind of selection to it to recover a population mostly in the OFF state. The cycle is then completed by applying a second form of selection that returns the population to a mostly ON state ready for another cycle.

1.7.1 Phase variable loci *cj1421c* and *cj1422c* code for phosphoramidate transferases

C. jejuni has an unusual form of surface modification in the addition of O-methyl phosphoramidate (MeO*P*N) groups to the CPS. MeO*P*N is a phosphate group in which one oxygen is replaced by an amine group and a methyl group is attached to a second oxygen. As such it is a large and highly negatively charged modification.

In the NCTC 11168 strain this group can be attached in two positions on the CPS, the first is attachment to the Gal*f*NAc group, and the second is to the Hep sidechain (McNally *et al.*, 2007) as shown in figure 1.5. Both of these attachments are controlled by phase variable genes: *cj1421c* and *cj1422c* respectively, with the ON state producing attachment and the OFF state resulting in an OH group in the same place.

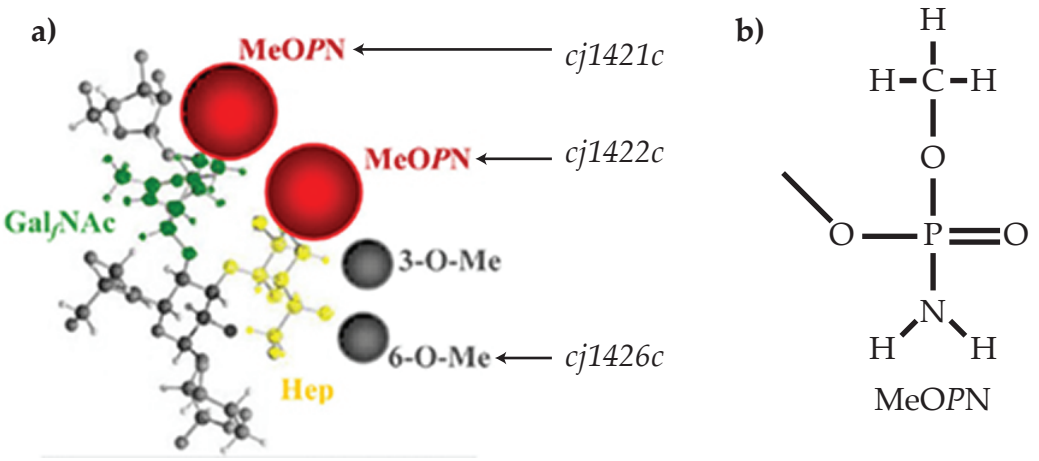

Figure 1.5: Attachment sites of the MeO*P*N groups on the CPS
The position of MeO*P*N attachment on the CPS is shown in (a). The gene names to the right indicate three phase variable genes that affect attachment of moieties to this part of the CPS. Cj1421c attaches an MeO*P*N group to the Gal*f*NAc group, whilst Cj1422c attaches an MeO*P*N group to the Hep side chain. Cj1426c, which is also phase variable (PV), attaches a nearby 6-O-Methyl group. The structure of the MeO*P*N is shown on the right (b).

1.7.2 Phage F336 resistance in NCTC 11168 is determined by presence or absence of capsular MeO*P*N groups

Phage F336 is an obligate lytic *C. jejuni* bacteriophage isolated from duck intestinal contents in Denmark. It is classed as a category II phage from the *Myoviridae* family (Hansen *et al.*, 2007). Further experiments with this phage showed that the NCTC 11168 strain of *C. jejuni* was able to rapidly develop resistance to this phage and a resistant strain was isolated (Sørensen *et al.*, 2011). Analysis of this resistant strain and of other resistant strains showed that the sensitivity of the strain is critically dependent on the state of the two phase variable genes *cj1421c*

and *cj1422c*. Specifically, it showed that if *cj1421c* is ON and *cj1422c* is OFF then the isolate will be sensitive, and all other states are resistant (Sørensen *et al.*, 2011; 2012). Thus, the organism can move to a resistant state either by switching *cj1421c* OFF or switching *cj1422c* ON. The mechanism behind this resistance is currently unknown, but it is hypothesized that the MeO*P*N group attached by *cj1421c* forms part of the phage attachment binding site while the MeO*P*N group attached by *cj1422c* blocks that binding site (Sørensen *et al.*, 2012). The same study also showed that the 6-O-Methyl group attached by another phase variable gene, *cj1426c* also impacts phage sensitivity but to a lesser degree since it roughly halves the efficiency of plating (EOP) rather than reducing it to zero, or nearly zero (Sørensen *et al.*, 2012).

1.7.3 Other roles of the MeO*P*N group

Champion *et al.* (2010) argued that the MeO*P*N group has insecticidal properties, based on experiments in larval *Galleria mellonella* in which they demonstrated that knocking out *cj1416c* and thus the MeO*P*N biosynthesis pathway (McNally *et al.*, 2007) inhibits *G. mellonella* killing by *C. jejuni*. However, further experiments by van Alphen *et al.* (2014) demonstrated that although disrupting the biosynthesis pathway impacted survival, specifically removing the MeO*P*N transferases or directly injecting MeO*P*N modified CPS or MeO*P*N containing compounds into the insects had no effect on their survival. This suggests that another effect of modifying the MeO*P*N biosynthesis pathway is responsible for the changed survival rate of the larvae and not the MeO*P*N groups themselves.

van Alphen *et al.* (2014) also looked at a range of other influences of the MeO*P*N groups. Most of these experiments are carried out in the 81-176 strain of *C. jejuni* which is another common laboratory strain of *C. jejuni*. In 81-176, the two MeO*P*N transferases are *cjj81176_1420* and *cjj81176_1435*. It is not known whether these two genes attach the MeO*P*N moeity in the same locations as in NCTC 11168 but

both are phase variable. As with NCTC 11168 these genes have a large homology region shared between the two genes and this region contains the poly-G tract. This homology makes the genes prone to homologous recombination between the two sites and thus they investigated the function of the genes only with a double knockout strain. In addition to the insecticidal activity discussed above these experiments looked at the impact of MeO*P*N groups on invasion of epithelial cells (reduced by MeO*P*N), serum resistance (increased), chicken colonization (no effect) and piglet colonization (increased).

Of these, serum-mediating killing is of particular interest due to its potential to be used as a selective agent in the laboratory and because it is the ON state that is beneficial. The 81-176 strain is resistant to killing by human sera in its wildtype state, but van Alphen *et al*. (2014) showed that knocking out either the entire MeO*P*N biosynthesis pathway[3] or the two MeO*P*N transferases resulted in dramatically reduced survival in serum (figure 1.6). These experiments supported the findings of (Maue *et al*., 2013) who has previously found that disrupting the biosynthesis pathway influenced killing by human serum. The addition of EGTA almost completely inhibited serum-mediated killing suggesting that the killing operates via the classical complement pathway (Des Prez *et al*., 1975).

1.8 Structure and scope of this work

The goals of this project are: (1) to investigate the diversity and conservation of phase variable genes in *C. jejuni* and *Campylobacter* in general; (2) to develop a cyclical selection assay and use that assay to interogate questions relating to the biological significance of differences in phase variation rate; and (3) to create an *in*

[3]This was achieved by knocking out *cjj81176_1415* which is homologous to *cj1416* in NCTC 11168

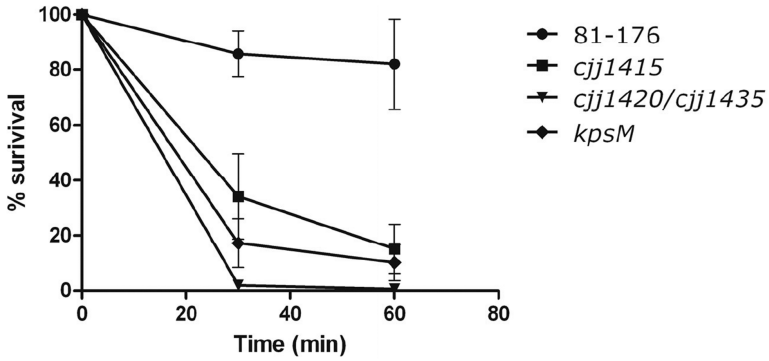

Figure 1.6: Knocking out MeO*P*N transferases increases serum sensitivity
Figure shows survival of four strains of *C. jejuni* 81-176 in 10% human serum over time. Line labelled 81-176 is the wild type, *cjj1415* is from a knockout strain of *cjj1415* which disables the MeO*P*N biosynthesis pathway, *cjj1420/cjj1435* is from a knockout strain removing the two MeO*P*N transferases, and *kpsM* is an uncapsulated knockout strain for comparison. Figure is reproduced from van Alphen *et al*. (2014) under the Creative Commons Attribution (CC BY) license 4.0.

silico model that can be used to predict, and understand, the results of experiments carried out using this cyclical selection assay.

I begin with a novel survey of the phase variable genes present in *Campylobacter* strains, followed by a search for associations between the host strains were isolated on and the phase variable genes present in these strains. This serves to place the work of the later chapters into the wider context of phase variation in *Campylobacter*. The rest of my work is concerned with the development and modelling of a cyclical selection assay that allows the cycle of selection and counter selection of a chosen gene to be studied under controlled conditions. Chapter 4 deals with the development and verification of the selection and counter-selection components of this cycle, while in chapter 5 I demonstrate the viability of the selective cycle and discuss its limitations and potential issues. Finally, in chapter 6 I develop an *in silico* model of cyclical selection and use the results of chapter 4 to predict the behaviour of the eventual cyclical assay. Additionally, I extend this model to study the behaviour of non-selective bottlenecks.

Chapter 2

Materials and Methods

2.1 Microbiological techniques

2.1.1 Normal bacterial growth conditions for *C. jejuni*

C. jejuni strains were grown at 42°C under microaerobic conditions (4% O_2, 10% CO_2, 86% N_2) maintained by a Thermo-Scientific VA1000 workstation. Strains cultured on solid media were grown on Mueller-Hinton Agar (Oxoid) which was routinely supplemented with 5µg/µl of Trimethoprim and 10 µg/µl of Vancomycin to reduce the chances of contamination. Liquid cultures were grown in Mueller-Hinton Broth (Oxoid) with shaking.

2.1.2 Alternative bacterial growth conditions for *C. jejuni*

Experiments described in section 4.5.1 were carried out in the laboratory of Dr Lone Brønsted at the University of Copenhagen, Denmark. For these experiments,

strains were grown at 37°C under microaerobic (6% O_2) conditions on Blood Agar plates (made using Base II (oxoid), supplemented with 5% defibrinated horse blood).

2.1.3 Bacterial growth conditions for *E. coli*

E. coli was grown at 37°C on lysogeny broth (LB) agar plates under normal aerobic conditions. Plates were supplemented with selective antibiotics as needed to maintain persistence of any plasmids present in strains. Liquid growth conditions were at 37°C in LB with shaking.

2.1.4 Long term storage at -80°C

C. jejuni isolates were stored by suspending overnight cultures grown on solid media in Mueller-Hinton broth (MHB) and then mixing 500µl of this suspension with an equal volume of 50% glycerol for a final concentration of 25% glycerol. *E. coli* strains were stored by growing overnight liquid cultures in LB and then mixing 500µl of this culture with an equal volume of 50% glycerol for a final concentration of 25% glycerol. These were then placed at -80°C for long term storage without flash freezing. Strains were extracted by scraping out with a sterile loop and streaking onto suitable solid growth media.

2.2 Bacterial strains used

The following strains were used in this thesis:

Species	Strain	Obtained from
C. jejuni	3447 (MP19)	Dr Lone Brønsted[1]
C. jejuni	NCTC 11168	Dr Richard Haigh[2]
C. jejuni	NCTC 11168 rpsL*	Dr Richard Haigh[2]
C. jejuni	NCTC 11168 Δ*cj1422c*::RDH315	This project
C. jejuni	NCTC 11168 MP21	Dr Lone Brønsted[1]
E. coli	DH5α	Dr Richard Haigh[2]

[1] Department of Veterinary Disease Biology, University of Copenhagen

[2] Department of Genetics, University of Leicester

2.3 28-locus-CJ11168 PV-analysis assay

In order to identify both the overall state of PV loci and the individual phasotype composition, fragment analysis was used to identify the state of 28 phase variable loci of NCTC 11168 using the method described in Lango-Scholey *et al.* (2016). This assay worked by polymerase chain reaction (PCR) amplification of fragments containing the target SSRs using primers labelled with fluorescent tags. Detection of these tags by capillary electrophoresis allows the exact size of each fragment to be determined and thus the number of repeats in the SSR to be determined. Primer pairs were designed to ensure amplified fragments are distinguishable by either size or dye type so that 28 loci could be measured in a single capillary electrophoresis assay.

Analysis was carried out in 96-well PCR plates. Samples were prepared from a colony by picking the colony from the plate and re-suspending in 100µl of purified distilled water (ELGA). Cells were then lysed by heating to 98°C for 5-8

minutes followed by centrifugation at 768×G for 4 minutes to pellet cell debris. The supernatant was then collected for further analysis.

Samples were processed by PCR amplification (using cycle shown below) with six sets of multi-plexed primers (listed in appendix A). 2µl of each PCR reaction was mixed with fresh *Taq* polymerase and additional buffer, dH_2O, and $MgCl_2$ to maintain a constant solution and heated to 72°C for 45 minutes to allow the attachment of untemplated A-residues by the *Taq* polymerase to complete. 0.5µl of this mix was then diluted in 9.25µl of foramamide and 0.25µl of GeneScan™ LIZ-500 or LIZ-600 size standard. This was then passed to The Protein and Nucleic Acids Chemistry Lab (PNACL) for capillary electrophoresis using an Applied Biosystems 3730 Genetic Analyser.

Temperature	Time	
94°C	5 minutes	
98°C	30 seconds	×25
50°C	30 seconds	
72°C	60 seconds	

Table 2.1: PCR conditions for 28-locus-CJ11168 PV-analysis assay

The data returned from electrophoresis was analysed by Peak Scanner™ to produce a list of peaks present in the electrophoresis data which can be exported into a tab-separated file. This file can then be automatically analysed using a Perl script (PSAnalyse, see section 4.3) to calculate fragment lengths by comparison to a sample of known tract lengths (as confirmed by sequencing) and scored as ON or OFF based on whether the tract length produces the longest possible reading frame. In the case of four tracts which are either intergenic or located in pseudogenes an arbitrary ON length was chosen (see table 1.1).

2.4 Molecular Genetics

2.4.1 Digestion with restriction enzymes

Digestion with restriction enzymes was carried out using enzymes from New England BioLabs Inc. (NEB) using the mix specified below for a 25µl digestion. Digestions were left for 1 hour at the temperature specified by the manufacturer. Choice of which of four buffers to use in double digestions was determined by use of NEB's online Double Digest Finder tool[1]. Bovine serum albumin (BSA) was pre-prepared at 10× concentration and added if specified by NEB.

Component	Single	Double
DNA	∼0.5µg	∼0.5µg
10× buffer[1]	2.5µl	2.5µl
10× BSA[2]	2.5µl	2.5µl
Enzyme 1	0.5µl	0.5µl
Enzyme 2	–	0.5µl
dH_2O[3]	to 25µl	to 25µl

[1]Buffer 1-4 as specified by NEB

[2]If needed

[3]ElgaStat option 2 purified distilled water (ELGA)

2.4.2 Ligation with T4 DNA ligase

Ligation was performed with T4 DNA ligase (NEB) according to manufacturer's instructions in 10µl volumes. For insertion of a single insert into a vector a 3:1

[1]NEB have since replaced this finder with a new version for their current buffer system.

molecular ratio of insert to vector was used; for the insertion of two fragments into a vector, a 5:5:1 molecular ratio was used; and for three way ligation of parts, a 1:1:1 molecular ratio was used. Total DNA concentration was kept in the range of 1-10µg/ml. Reaction was allowed to proceed overnight at 16°C or for 3 hours at 37°C. After ligation product was purified using an E.Z.N.A Cycle Pure kit (Omega bio-tek) according to manufacturer's instructions.

Component	Volume
10× T4 ligase buffer	1µl
DNA for ligation	see text
T4 ligase	0.5µl
dh_{20}[1]	to 10µl

[1] ElgaStat option 2 purified distilled water (ELGA)

2.4.3 Preparation of chemically competent *E. coli* strains

Chemically competent *E. coli* strain DH5α cells were prepared from cultures grown in 5ml LB overnight at 37°C with shaking. 1ml of this overnight culture was transferred into 100ml of fresh media and grown at 37°C with shaking until the OD_{600} reached roughly 0.4. This culture was then divided into two 50ml Falcon tubes, chilled on ice for five minutes and then pelleted by centrifugation for 15 minutes at a relative centrifugal force (RCF) of 2000×g at 4°C. The supernatant was discarded and both pellets were re-suspended in 20ml of filter sterilised TFB1 (30mM KOAC, 50mM $MnCl_2$, 100mM KCl, 10mM $CaCl_2$, 15% v/v glycerol). This suspension was centrifuged to re-pellet the cells – 5 minutes at RCF of 2000×g at 4°C – and the supernatant discarded. The pellets were then re-suspended in 4ml of filter sterilised TFB2 (10mM Na-MOPS (pH 7.0), 10mM KCl, 75mM $CaCl_2$, 15% v/v glycerol). This produced chemically competent cells which were then aliquoted into 1.5ml tubes and could be stored at -80°C until needed.

2.4.4 Heat-shock transformation of *E. coli*

Chemically competent *E. coli* cells were transformed by heat shock. 5-15µl of plasmid/DNA preparation was mixed with 50µl of chemically competent cells on ice and left to incubate for 30 minutes. These cells were then heat-shocked by heating to 37°C for 5 minutes using a temperature-controlled water bath. After heat shock cells were placed back on ice for 5 minutes before being mixed with 500µl of LB and incubated for 90 minutes at 37°C with shaking to allow time for the plasmid to express any antibiotic resistance. Cells were then plated onto lysogeny broth agar (LA) plates containing a suitable selective antibiotic at two densities (50µl and 500µl). Plates were then checked after 24-36 hours for resistant colonies. The presence of the plasmid in selected colonies was checked by colony PCR and positive colonies subcultured for further use and analysis.

2.4.5 Design of inserts for transformation of *C. jejuni* by homologous recombination

Correctly designed suicide plasmids inserted into *C. jejuni* will naturally insert target DNA sequences into the chromosomal DNA by homologous recombination. This requires 400bp or more, and ideally 500bp or more, of homologous sequence upstream and downstream of the target insertion site. The removed sequence can be several thousand base pairs in length as can the inserted DNA. Including an antibiotic resistance cassette in the insert allows the recombinant strain to be selected for after transformation. The suicide plasmid carrying this insert must be suitable for replication in *E. coli* to allow a high concentration plasmid preparation to be created and may carry additional selection markers for use in *E. coli*.

2.4.6 Transformation of *C. jejuni* by electroporation

C. jejuni cells for transformation were prepared by plating at high density (for confluent growth) onto MHA plates, and incubating for 24-36 hours under normal growth conditions. 2ml of chilled wash buffer (272mM sucrose and 15% v/v glycerol, filter sterilised) was added to each plate, and bacterial cells scraped into suspension with a sterile spreader. Liquid was transferred into 1.5ml microcentrifuge tubes and then centrifuged in a benchtop microcentrifuge (Heraeus Biofuge pico) at maximum speed for 3 minutes to pellet the cells. The supernatant was discarded and the cells re-suspended in 1ml of ice cold wash buffer. This washing step was repeated 3 further times. After the final wash, the cells were re-suspended in 100µl of wash buffer.

50µl of prepared cells was mixed with 1-5µg of plasmid DNA in distilled water for transformation in pre-chilled electroporation cuvettes (2mm gap). The cells were pulsed once using a BioRad Gene Pulser and BioRad Pulse Controller (Biorad, UK) with settings of 2.5kV, 200Ω, and 25µF. Following electroporation the time constant was checked for a resulting value of approximately 4.8. 200µl of MHB was added to the cells immediately before spreading onto two non-selective MHA plates, 10% of the mixture onto the first plate and the remainder onto the second plate. Plates were incubated for 5 hours under normal *C. jejuni* growth conditions and then harvested into 1ml MHB and spread onto fresh MHA plates containing selective antibiotic. Colonies were collected after 3-5 days and checked by colony PCR and sequencing to ensure correct insertion.

2.4.7 Use of rpsL* mutant to produce markerless mutants

The *C. jejuni* NCTC 11168 rpsL* strain carries a single point mutation to the *rpsL* gene (*rpsL**) that renders the strain resistant to Streptomycin (Haigh, personal

communication). The RDH315 insert from the pRDH315 plasmid carries a Chloramphenicol resistance cassette (*cat*) together with *hprpsL* which is the *rpsL* gene from *Helicobacter pylori*. The inclusion of *hprpsL* is intended to provide the basis for counter-selection against the cassette. RpsL forms a dimeric complex, so the inclusion of a wild type *rpsL* produces heterodimer with the protein produced by *rpsL** that can be targeted by Streptomycin and thus render the strain sensitive to the antibiotic. The *H. pylori* version of the gene is used instead of the *C. jejuni* version to limit the spontaneous production of a resistant mutant by homologous recombination between the inserted *rpsL* gene and the native *rpsL** gene. The removal of the insert with a second suicide plasmid can then be selected for by incubation with Streptomycin. This selection and counter-selection technique is illustrated in figure 2.1 but see section 4.4 for problems with the deployment of this technique.

The *C. jejuni* NCTC 11168 rpsl* strain and the pRDH315 plasmid were kindly provided by Dr Richard Haigh (University of Leicester).

2.5 Phage propagation/manipulation

Phage F336 was kindly provided by Dr Lone Brønsted (Department of Veterinary Disease Biology, University of Copenhagen), as was *C. jejuni* strain 1447 (Hansen *et al.*, 2007) used for phage propagation.

2.5.1 Media and buffers for phage use

SM buffer with gelatin (Cold Spring Habour Protocols, 2006) was used as a storage buffer for bacteriophage. SM buffer with gelatin was prepared with 50mM Tris-Cl,

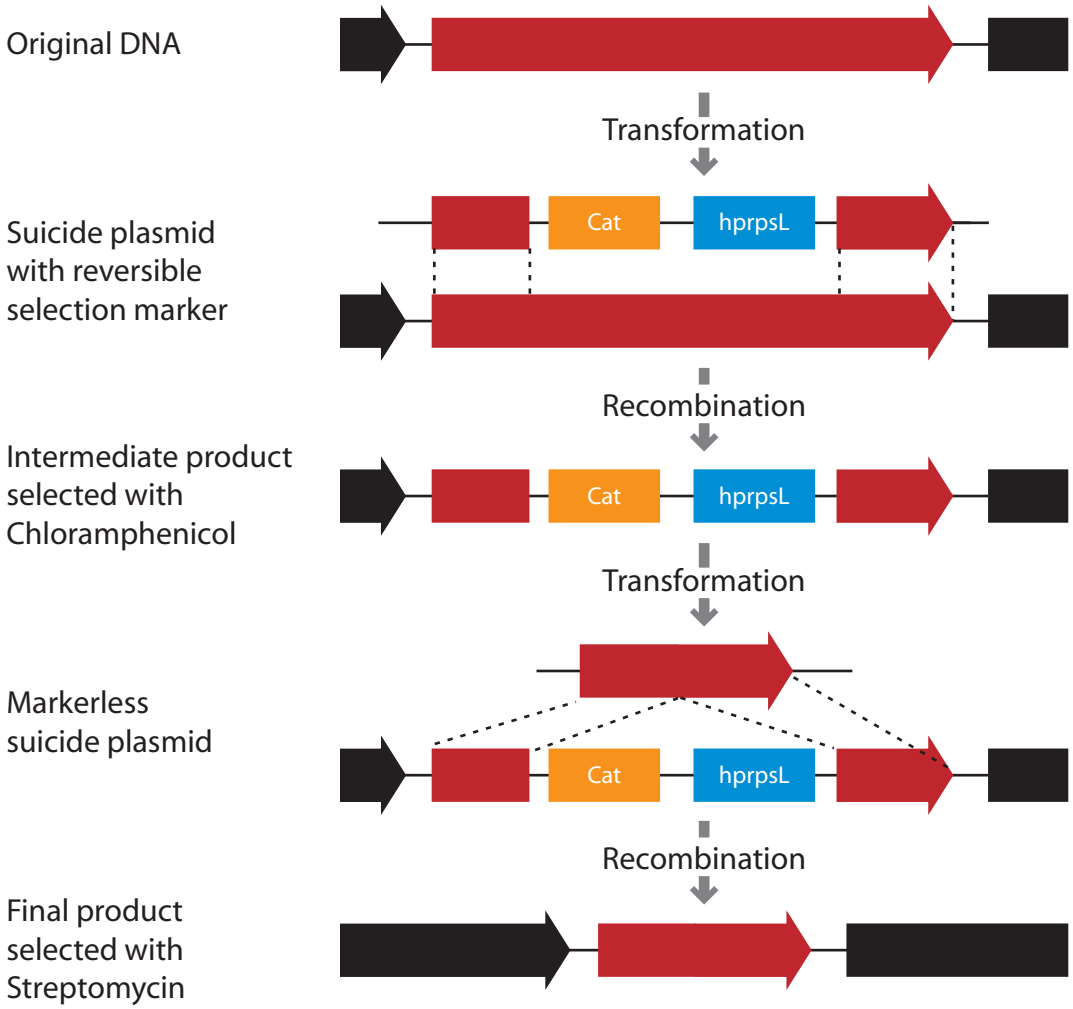

Figure 2.1: Creation of markerless mutant using rpsL counter-selection
Diagram illustrates method of markerless insertion. Initially, a suicide plasmid containing selection (*cat*) and counter-selection (*hprpsL*) genes is inserted into Streptomycin resistant rpsL* mutant strain of *C. jejuni* by suicide plasmid with homologous flanking regions to target the insert to the desired site. Homologous recombination produces a Streptomycin-sensitive mutant with both selection and counter-selection markers that can be selected using Chloramphenicol resistance. A second suicide plasmid containing only the flanking region (or the flanking region and any targeted mutational changes) is inserted into the new mutant. Homologous recombination again modifies the chromosomal DNA, and the removal of the counter-selection marker returns the strain to being Streptomycin resistant allowing the markerless mutant to be selected with Streptomycin.

pH 7.5, with 100mM NaCl, 8mM $MgSO_4$, 0.01% w/v gelatin, and made up to volume with purified distilled water (ELGA, ElgaStat Option 2). Cation-enriched brain-heart infusion (CBHI) media was made by making brain-heart infusion (BHI) broth (Oxoid) according to manufacturers instructions and then supplementing with $CaCl_2$ and $MgSO_4$ to final concentrations of 1mM and 10mM, respectively. Supplementation with $CaCl_2$ and $MgSO_4$ was carried out after sterilisation by autoclaving and so media was left overnight at 37°C after supplementation to allow any contamination to be detected.

NZCYM broth (Sigma-Aldritch) was used to produce top and bottom agar for phage titration and propagation. In both cases, NZCYM broth was made to manufacturers instructions using purified distilled water (ELGA, ElgaStat Option 2) and then supplemented with BioAgar (BioGene) to produce solid media: 0.6% w/v for top agar, and 1.2% w/v for bottom agar. Bottom agar plates were supplemented with 10µg/ml vancomycin to reduce the potential for contamination.

2.5.2 Phage propagation

Phage propagation was carried out as described in Sørensen *et al.* (2011). *C. jejuni* strain 1447 was cultured overnight at high density on blood agar or MHA plates before scraping into CBHI. This suspension was normalised to an OD_{600} of 0.35 and then incubated under normal *C. jejuni* growth conditions (section 2.1 above) for 4 hours before mixing with bacteriophage in SM buffer with gelatin at an approximate multiplicity of infection (MOI) of 0.01. The mixture was left for 15 minutes to allow for phage absorption before mixing 0.6ml of suspensions with 5ml of top agar tempered to 45°C and then spread on pre-set bottom agar plates. After solidification the plates were incubated overnight under normal *C. jejuni* growth conditions before flooding with 5ml SM buffer and placing overnight at 4°C under aerobic conditions with gentle shaking to allow the phage to diffuse into

the SM buffer. Finally, the buffer was collected and passed through a sterile 0.2-µm filter to separate the phage from any bacteria that had also diffused into the buffer. The concentration of phage in plaque forming units (PFU)/ml of the bacteriophage preparation was determined by titration (below). This bacteriophage suspension was stored at 4°C for a maximum time of 3 months before use.

2.5.3 Phage titration

Phage titration was carried out by plaque assay as described in Sørensen *et al*. (2011). NZCYM overlay plates were prepared as described above but without the addition of bacteriophage to the bacterial suspension introduced into the top agar. After solidification plates were left to dry for 45 minutes before serial dilutions from 10^0 or 10^{-2} to 10^{-7} were spotted onto the media in either 3 times 10µl or 6 times 5µl spots. Plaques were allowed to dry at room temperature before plates were transferred to normal *C. jejuni* growth conditions for overnight growth. Following incubation, plaques were counted and the PFU/ml of the undiluted suspension calculated from these counts.

2.6 Selective protocols introduced in this thesis

2.6.1 Phage selection assay

C. jejuni was grown on blood agar or MHA plates overnight then scraped into CBHI. This suspension was adjusted to a final OD_{600} of 0.01 ($\sim 10^7$ colony forming units (CFU)/ml) and a volume of 20ml in a sterile petri dish and then incubated under normal growth conditions for 4 hours. A small sample was taken to allow exact CFU counts to be determined by serial dilution and individual colonies to be

picked for phasotype analysis. After incubation phage F336 was added to the *C. jejuni* culture at an MOI of approximately 0.001 ($\sim 10^4$ PFU/ml) and incubated for a further 24 hours under normal growth conditions. After incubation a sample was serially diluted onto plates and incubated under normal growth conditions to allow for individual colony collection after 3 days.

2.6.2 Serum selection assay

C. jejuni was grown overnight in Mueller-Hinton broth, pelleted by centrifugation and then re-suspended in 1ml HEPES buffer (10mM HEPES, 150mM NaCl, 5mM $CaCl_2$, 5mM $MgCl_2$, pH 7.4) and then normalised to an OD_{600} of 0.2. Serum selection was carried out by adding mixing with pooled human serum from healthy volunteers to the desired final serum concentration (10%, 25%, or 50%) in a total volume of 500µl and incubating at 37°C for one hour. For some experiments 50µl samples were taken during incubation at 0 minutes and 30 minutes. Samples taken during the incubation and at the end were serially diluted onto MHA plates to collect single colonies for the 28-locus-CJ11168 PV-analysis assay analysis and to determine viable cell counts.

2.7 Computational methods

2.7.1 *in silico* simulation

in silico simulation was performed using software written in Python (version 3.3.5 and 3.5.0, available from python.org) with random number generation by numpy (versions 1.8.0 and 1.8.2, Van Der Walt *et al*. 2011, available from numpy.org).

2.7.2 PSAnalyse

PSAnalyse (see section 4.3 was written in Perl 5 (interpreted by StrawberryPerl version 5.16.3.1), with a frontend written in C# 4.0 using Visual Studio 2015 community edition. Packaging of the Perl script into an executable for redistribution was carried out using Cava Packager 2.0.

2.7.3 PhasomeIt

Phasome*It* (see chapter 3) was written to be Python 3.3 compatible and interpreted with Python 3.5.0 (available from python.org), with the packages BioPython version 1.65 (Cock *et al.*, 2009) and Natsort version 5.0.1 installed. BioPython was used to access the Entrez gateway to the National Center for Biotechnology Information (NCBI) databases to acquire genome sequences. The output is generated in HTML 5 and uses JavaScript to dynamically alter pages. The sorttable[2] package is included as part of the output and is used under the MIT license[3]. Bossref (section 3.3) was written in C++11 and compiled using either Visual Studio 2015 community edition (for Windows) or GCC version 4.8.4 (for Linux).

2.7.4 Production of figures

Graphs were produced using R 3.2.4 ("Very secure dishes", 64-bit) (R Core Team, 2016) with packages GGPlot2 (Wickham, 2009), gridExtra (Murrell, 2005), and plyR (Wickham, 2011). Other figures were produced using the LaTeX package TikZ (Tantau, 2013), Adobe Illustrator CS2, or exported from the software described

[2] http://www.kryogenix.org/code/browser/sorttable/
[3] http://www.kryogenix.org/code/browser/licence.html

herein. Some images were annotated with Adobe Photoshop CS2. Figures were checked for suitability for a colour-blind audience with Color Oracle (Jenny and Kelso, 2007).

2.8 Statistical analysis

2.8.1 Significance testing of clustering within trees

Clustering within trees was tested using a tree-based scanning method with conditional Poisson model (Kulldorff *et al.*, 2003). This analysis was carried out using TreeScan 1.3 (Kulldorff, available at `http://www.treescan.org`).

2.8.2 Statistical testing of proportions ON and OFF

Differences in the frequency of OFF and ON phase states in each gene under the phage present and control conditions were fitted to a generalised linear mixed model using a logit link function. Calculations were carried out in R 3.2.5 (R Core Team, 2016), models fitted and p values obtained from Wald's Z using the 'glmer' command (family = 'binomial') in the lme4 package (Bates *et al.*, 2015) and graphs plotted using ggplot2 (Wickham, 2009). The conservative Bonferroni correction was used to account for multiple hypothesis testing where appropriate.

2.8.3 Approximation of 95% confidence intervals for proportions ON

95% confidence intervals for true proportions were calculated using the Mid-P adaptation of the Clopper-Pearson interval (Agresti and Gottard, 2005) which

gives a closer approximation of a true 95% confidence interval than conservative methods. Calculations were carried out in R 3.2.5 (R Core Team, 2016) using the PropCIs package (Scherer, 2014).

Chapter 3

The phase variable genes present in *Campylobacter* are highly variable between strains

3.1 Summary

In this chapter I investigate the diversity of phase variable genes present in different species and isolates of *Campylobacter*. The phasome of an isolate is defined as the set of phase variable genes of that isolate. Already published work indicates that there is considerable variation in the number of phase variable genes in the phasome of *C. jejuni* isolates but no wider comparison or survey of the *Campylobacter* phasome exists. Here I present a new program to analyse the phasome from genomic data as well as an analysis of the data produced by running this program over two genome collections: (1) 77 complete *Campylobacter* genomes acquired from NCBI, and (2) a

set of 190 incomplete *C. jejuni* and *C. coli* genomes with known source attribution and metadata provided by Professor Samuel K. Shepperd (University of Bath). The analysis of these datasets reveals surprising diversity in the phasomes of different isolates with a huge number of rare phase variable genes with homologues present in only a handful of – or in many cases: one – isolates as well as a small number of broadly conserved phase variable genes present in the majority of strains and a core phasome common to strains from the same species. Analysis of the dataset with known host attribution indicates that there are some homology groups of phase variable genes which are particularly associated with isolates obtained from cattle in one ST complex.

3.2 Introduction

When the first *Campylobacter* genome was published (Parkhill *et al.*, 2000) the presence of a large number of highly variable poly-G tracts was noted. Since then, as more genomes have been sequenced (e.g. Pearson, 2013, Kienesberger *et al.*, 2014), the presence of these tracts has been discovered in other strains and species. However, while some comparisons of presence and absence have been made (e.g. Bayliss *et al.*, 2012) no widescale survey of the phase variable genes of *Campylobacter* has been performed. The work in this chapter seeks to address this absence. In order to ease discussion, it is useful to have a term that can be used to refer to the collection of phase variable genes present in an isolate, strain or species. I have therefore coined the term 'phasome' to refer to these sets and will use this term throughout this thesis.

Phase variation has been identified as involved in a wide variety of important aspects of bacterial biology (see section 1.2 and section 1.5) so characterisation of the phasome may reveal important features about the ecology and evolution of these

bacteria. The falling cost of sequencing has changed the major problem of genomic data from acquiring it in the first place to analysing the data into understandable forms. In this analysis, I concentrate on phase variation generated by variation in SSRs which are the major generator of phase variation in *Campylobacter* although other mechanisms do operate in some species (see section 1.5.1). Identification of SSRs can be achieved by a search with pattern recognition, but the identification of PV repeats requires a cutoff to be chosen between shorter non-PV and longer PV tracts. This has been taken as 7Gs or longer in the existing literature (Pearson *et al.*, 2007, Bayliss *et al.*, 2012). Once the SSRs have been identified they can be associated with particular genes by analysing the annotation present on the genome.

Lists of PV genes present in each isolate from a large collection are of limited value for comparison between strains and species. To do this information is needed on the similarity between the different phasomes. Pairwise comparisons are likely to be prohibitively slow for large datasets, would both be reliant on the quality of the annotation, and are limited to comparing PV genes and thus exclude any similarities to non-PV genes. In contrast, a genome wide search for homologies does not suffer from these limitations. Although other alternatives now exist (e.g Kielbasa *et al.*, 2011, Huson and Xie, 2014), the venerable Basic Local Alignment Search Tool (BLAST) suite (Altschul *et al.*, 1990, Camacho *et al.*, 2009) continues to be a widely used and highly effective tool for identifying homologous sequences and so it is this suite of tools that were chosen for use.

The BLAST algorithm looks for sequences similar to a query sequence within a set of subject sequences. These sequences can be either DNA or amino acid sequences, or translated amino acid sequences generated from the DNA sequence. The first step is to process the subject sequence(s) into a tree structure to allow efficient searching; this processed structure is referred to as a database. This tree structure is used to identify short sequences (k-mers) that match parts of the query sequence. The algorithm then searches the flanking sequences of these short sequences to

extend the match in both directions. A heuristic scoring mechanism is used to assess the quality of the match, applying a penalty to mismatches and gaps (and, for amino acid sequences, applying a lower penalty for chemically similar residues). The scoring heuristic can be chosen from a range of theoretically derived scoring matrices (e.g. PAM-30, BLOSUM-62, etc.) with the choice depending on the length of query sequence and the evolutionary distance between subject and query sequence (Pearson, 2013).

3.3 Use of Exclusive OR for highly efficient and customisable identification of SSRs

The detection of repeats by a simple text search is practical for poly-G and poly-C tracts but because the number of resultant combinations rapidly increases when searching for longer tract unit lengths this approach rapidly becomes inefficient. Although there are multiple existing programs able to identify SSRs (reviewed in Merkel and Gemmell, 2008), none of these provide the required fine control over which sequences are identified. In particular, none offer control over length cutoffs specific to each repeat length and many are designed to find imperfect repeats that are not relevant to phase variation. Therefore I developed a program called Bossref (for BOolean Simple Sequence REpeat Finder) to identify these sequences.

Bossref uses a novel algorithm which relies on the behaviour of the simple Boolean operator 'xor' (eXclusive OR). Xor takes two binary inputs (bits) and returns 1 if the two bits are different and 0 if they are identical – i.e. "either A or B but not both". By first converting the nucleotide sequence into a binary format with 2 bits for each base, and then xoring the sequence with its left shifted equivalent, SSRs can be rapidly identified by counting the number of sequential 0s produced (see figure 3.1 for an illustrative example). Since each base is represented by 2 bits, the

degree of left shifting required is equal to twice the length of repeat unit to be detected. This also limits the maximum detectable repeat to one quarter the bit width of the integer representation used. With current 64-bit processors this limits the maximum detectable repeat unit size to 16 base pairs for the native integer size. This limit does not impact this work but could be extended to 32 base pairs by using an 128-bit integer type if needed. Although this process needs to be repeated for each length of repeat unit searched for it is linear in algorithmic complexity with respect to both sequence length and the maximum length of repeat unit to search for. This is an improvement over word-searching algorithms such as those used in TROLL (Castelo *et al.*, 2002) which while linear in algorithmic complexity with respect to sequence length are quadratic with respect to maximum length of repeat unit to detect. Moreover, because Bossref relies on basic Boolean operations it is amenable to extremely efficient implementation and thus significantly outperforms existing SSR search algorithms (table 3.1)

Sequence	. A T C G G G G G G G G C G
In binary	..0010011111111111111110111
Left shifted	0010011111111111111110111..
Exclusive OR	..1011100000000000000001010..

SSR produces run of 0s

Figure 3.1: Example of detection of SSRs by left shift and exclusive OR
In the example above, first the DNA sequence is converted to a binary code (2 bits per base) then the sequence is left shifted by twice the repeat unit length. Next it is xored with the original sequence which produces a run of 0s equal to the twice the repeat length less one. Detection of these runs of 0s allows the repeats to be identified. Increasing the magnitude of the left shift allows repeats consisting of longer repeat units to be identified.

Program	Time
Bossref	5.9s
TROLL	19.9s

Table 3.1: Speed comparison for Bossref
Speed comparison of Bossref and TROLL (Castelo *et al.*, 2002). Timings based on search for all perfect SSRs of 20bp or greater and repeat unit size of 1-5bp in the *Arabidopsis thaliana* genome. Performed on BioLinux 8.0.7 running in Oracle VirtualBox 5.1.2 with 4Gb of RAM on an Intel i5-6600K processor running at 4.2Ghz. Timings are the average of 5 runs of each program.

Bossref can be used separately from the main suite and has been made available on-line at http://www.jackaidley.co.uk/bossref. It takes as its input a FASTA formatted file containing one or more sequences and a list of cutoff levels for the minimum number of repeats of units of different repeat size together with an optional filter allowing particular types of repeat to be fine-tuned in their minimum repeat number. The program outputs a tab-separated output file giving information on location within the sequence, repeat unit type, and number of repeats for every SSR found. This output is designed to be machine readable for further analysis.

3.4 Phasome*It* will identify and analyse the phasome of medium sized genome sets

Before moving on to the phasome analysis, I will first describe the program developed to carry out the analysis, dubbed Phasome*It*. Phasome*It* has been made available online at http://www.jackaidley.co.uk/phasomeit.

3.4.1 Overview of analysis process

Phasome*It* takes as its target a collection of genomes in GenBank flat file format (Benson *et al.*, 2005). It begins by extracting FASTA versions of the sequences from these files and then uses Bossref (see section 3.3 above) to identify potentially variable SSRs. The cutoff levels for repeats are configurable to the target organisms, but for the *Campylobacter* sequences analysed in this chapter, the cutoffs were 7 or more Gs or Cs, 10 or more As or Ts, 6 or more dinucleotide repeats, and 5 or more tetra- and penta- nucleotide repeats. Trinucleotide repeats were excluded since they cannot produce frameshift mutations. The cutoff of 7 or more Gs or Cs

was chosen to match the cutoff used in the existing literature Pearson *et al.* (2007), Bayliss *et al.* (2012). A/T tracts are thought to have lower mutability than G/C tracts (Dornberger *et al.*, 1999, Bayliss *et al.*, 2004), so a higher cutoff for runs of As or Ts was applied. The cutoff of 10 As or Ts was chosen because tracts of this length appear to have similar levels of mutability to 7 Gs based on deep sequencing data (Bayliss group, unpublished data). Dinucleotide and longer tracts have been comparatively unstudied in *Campylobacter*, since they do not occur in the most studied species, so these cutoffs were chosen based on data from other species, including the closely related *Helicobacter pylori* (Saunders *et al.*, 1998).

Having identified the location of all SSRs that may result in phase variation, the program then uses the annotation data from each genome to identify the ORF associated with each SSR. When located within an ORF, the tract is associated with that ORF. When the SSR is not located within any ORF the program associates it with a nearby ORF. This is done by finding the closest ORF to a tract but with a bias for repeat tracts in the 5′ promoter region of a gene as these can influence expression of that gene. The program first finds the distance to the closest ORFs 3′ and 5′ of the SSR and then adds a biasing factor of 200bp if the tract is downstream of an ORF in the reading direction of that ORF. The tract-to-ORF distances are then sorted and the closest ORF is associated with the repeat tract. This biasing factor means that it is more likely to associate a tract with an ORF that the SSR is upstream of in the reading direction of that ORF.

Additionally, if the tract lies outside any annotated ORF the program looks for any potential reading frames missing from the annotation. It does this by finding the longest ORF starting from an ATG codon that spans the tract in any of the six possible reading frames including the possibility of a change in frame induced by a change in the length of the SSR. A minimum ORF length of 300 amino acids is then applied which exceeds the median length of annotated proteins in the dataset, and that of other species (Brocchieri and Karlin, 2005) and thus ensures a conservative

approach to identifying novel ORFs. If there exists such a reading frame then the tract will be associated with this rather than the nearest pre-annotated ORF.

The next step is to identify homologues of all of these genes in other strains. To do this a BLAST search is employed. To avoid over-dependence on the quality of annotation, a translated nucleotide search is used (tblastn) on the DNA sequence directly rather than relying on annotated protein sequences. For each ORF, the longest protein sequence generated by the three possible frames produced by varying the length of the SSR is used and BLAST-compared to every genome in the collection. Cutoffs of 50% subject coverage, 40% query coverage and an E value of 10^{-6} are used to filter the results. The BLOSUM-62 scoring matrix is used in these searches to allow for distant homologies.

3.4.2 Assignment of homology groups

PV genes are grouped into homology groups in a network fashion so that all genes that can potentially be connected by traversing a network of homology relationships are grouped together. This means that two PV genes can be in the same group even if they do not share a BLAST homology if both are homologous to a third gene. Note that although non-PV genes are also identified in the search and included in the homology groups further searches are not carried out to find homologues of these non-PV genes and thus they do not influence this network effect (see figure 3.2).

Homology groups are automatically assigned a name based on gene names and locus tags acquired from the annotation of PV genes within the group. Gene names are favoured over locus tags, and strains earlier in the ordering favoured over strains lower in the order. This ordering is primarily naturally sorted alphanumeric but Phasome*It* lets the user supply priority ordering so that species and strains about which more are known can be placed first in the order. This same order is

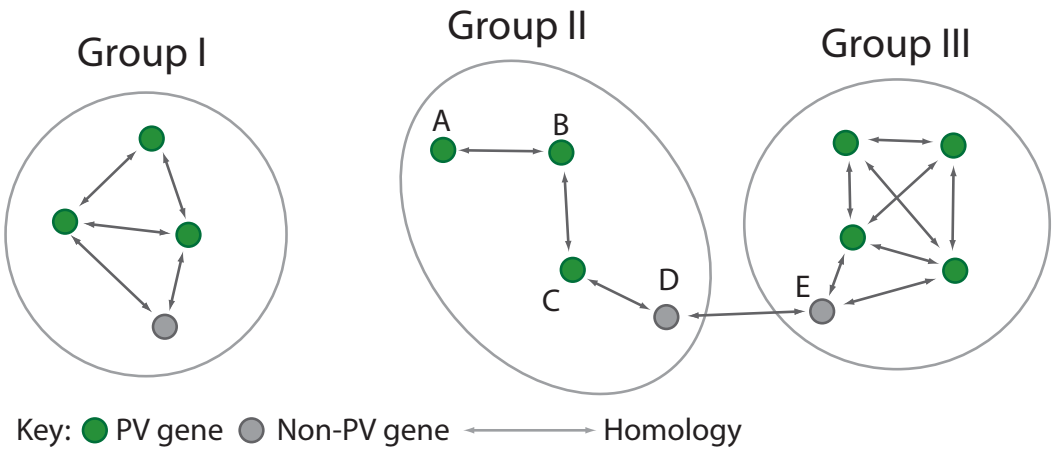

Figure 3.2: Grouping genes into homology groups
Coloured circles show identified genes, and arrows show homology relationships between genes. Genes that are connected by homology relationships are placed in the same group. Thus A and C are in the same homology group even though they share no homology relationship because they are both homologous to B. D and E do not connect groups II and III because D and E are non-PV homologues identified and the program does not search for homology relationships from these genes. Instead they are grouped only with PV genes homologous to them.

used to order the strains in other areas of the output. Together with the naming information the program also scrapes all the available functional annotation and applies a simple heuristic to extract the most relevant information. To do this it first discards low information words such as 'hypothetical' and 'protein' and then counts the number of times each word appears across all annotation data. The repetition of words between different annotations is used to assign a score to each functional annotation and the highest scoring annotation chosen as representative. The full list of annotations can also be viewed in the output.

3.4.3 Outputs from Phasome*It*

Phasome*It* generates a large number of HTML files that are viewable on any browser which fully supports HTML5 (use of Mozilla Firefox version 45 or greater is recommended). A central index file is provided that allows these files to be navigated by hyperlink and also provides some summary information. The primary outputs are (1) per-genome data, (2) tract length data, (3) gene grouping data, and (4) core phasome data. A file is produced for each genome listing all putatively PV SSR

located in the genome. For each SSR this file gives the position of that SSR within the genome, the type and length of SSR, a functional annotation, and any homologues. The tract length data provides summary and detailed breakdown data by repeat unit, species, and genome, as well as information on the putative ON lengths and a comparison of putative ON length and observed length.

Putative ON lengths are calculated by considering potential changes in tract length (i.e. an insertion or deletion of a single repeat) that would produce the two other reading frames and then comparing the length of the three potential translations before they reach a stop codon. The longest of these three protein products is considered to be the putative ON state of the protein, and the corresponding tract length to be the putative ON length. This prediction appears to be true in most cases, however it is known to be false for the *cj0170* tract in NCTC 11168 (see section 1.5.2) so cannot be taken as definitive.

The output of the gene grouping data begins with a summary graphic of the identified homology groups (figure 3.3 shows an example created with a reduced set of genomes). This is followed by a list of all homology groups and their members and then summary data on the rarity of the homology groups. Finally, trees based on the presence and absence of gene groups are created as well as alignment graphics for the identified BLAST matches. These trees can be coloured with information from supplied metadata or by species (e.g. figures 3.9 and 3.16).

Homology groups that are present in a large proportion of isolates of a species are referred to as the core phasome of that species. As a working definition, the core phasome of a species is defined as those homology groups that have a phase variable form of the gene in 60% or more of the genomes analysed from that species. In the output data, Phasome*It* identifies the core phasome in species with two or more representatives, and displays these with the identified group name and putative functional annotation, together with the proportion of genomes that they

are present in. To allow marginal cases to be examined, Phasome*It* applies a lower cutoff of homology groups present in 50% or more of the genomes from a species and displays these homology groups in addition to those in the core phasome.

3.5 The phasome of *Campylobacter* species

The first application of the program is to a collection of complete *Campylobacter* genomes obtained from the NCBI database (as of the date 5th September 2016). Some genomes were removed from this set: multiple sequences of different *C. jejuni* strain NCTC 11168 isolates (Revez *et al.*, 2012, Thomas *et al.*, 2014), and all but one sequence from a large ST-Complex 677 sequencing project that introduced 55 highly similar genomes to the database (Skarp *et al.*, 2015), were removed. These highly similar genomes were removed as they otherwise overwhelmed the signal from isolates represented by a single genome. Where multiple sequences for the same isolate were available all but the most recent was excluded. Finally, *C. fetus* strain 01/165 (accession no. CP014568) was removed as the sequence has annotation associated with it. This resulted in a final set of 77 genomes. These genomes were taken from 14 identified species with one further genome, RM16704, of an unidentified *Campylobacter* species. *C.jejuni* genomes made up the largest group (n=35), followed by *C. coli* (n=10), *C. fetus* (n=8), and *C. lari* (n=7). See table 3.2 for a complete listing of included isolates and accession numbers. For the main analysis, plasmids and other extra-chromosomal elements were removed and the analysis concentrated on the chromosomal genome alone. A separate analysis of plasmids associated with these strains was performed.

Important features of this analysis are shown on the gene group graphic (figure 3.3). This figure shows the output of a reduced input set, including only strains of *C. coli* and *C. fetus*, because the full figure is too large to reproduce on paper (it

Species	n	Strains (accession numbers)
C. coli	10	RM1875 (CP007183), RM4661 (CP007181), RM5611 (CP007179), FB1 (CP011015), BFR-CA-9557 (CP011777), HC2-48 (CP013034), OR12 (CP013733), CVM N29710 (CP004066), YH501 (CP015528), 15-537360 (CP006702)
C. concisus	2	13826 (CP000792), ATCC 33237 (CP012541)
C. curvus	1	525.92 (CP000767)
C. fetus	8	04/554 (CP008808), 97/608 (CP008810), pet-3 (CP009226), 84-112 (HG004426), 82-40 (CP000487), SP3 (CP010953), 03-427 (CP006833), cfvi03/293 (CP006999)
C. gracilis	1	ATCC 33236 (CP012196)
C. hominis	1	ATCC BAA-381 (CP000776)
C. hyointestinalis	2	LMG 9260 (CP015575), CCUG 27631 (CP015576)
C. iguaniorum	3	1485E (CP009043), 2463D (CP010995), RM11343 (CP015577)
C. insulaenigrae	1	NCTC 12927 (CP007770)
C. jejuni	35	R14 (CP005081), CG8421 (CP005388), MTVDSCj20 (CP008787), 00-2538 (CP006707), 00-2425 (CP006729), 00-2544 (CP006709), 00-2426 (CP006708), F38011 (CP006851), YH001 (CP010058), 00-1597 (CP010306), 00-6200 (CP010307), 01-1512 (CP010072), 00-0949 (CP010301), ICDCCJ07001 (CP002029), 35925B2 (CP010906), NCTC 11168 (AL111168), RM1221 (CP000025), 81-176 (CP000538), M1 (CP001900), S3 (CP001960), Isolate: IA3902 (CP001876), 269.97 (CP000768), 81116; NCTC 11828 (CP000814), 32488 (CP006006), RM3197 (CP012689), RM3196 (CP012690), CJM1cam (CP012149), NCTC11351 (LN831025), CJ677CC519 (CP010471), RM3194 (CP014344), OD267 (CP014744), WP2202 (CP014742), PT14 (CP003871), RM1285 (CP015209), 4031 (HG428754)
C. lari	7	LMG 11760 (CP007771), NCTC 11845 (CP007775), RM16701 (CP007777), RM16712 (CP007778), Slaughter Beach (CP011372), RM2100; ATCC BAA-1060D (CP000932), CCUG 22395 (CP007776)
C. peloridis	1	LMG 23910 (CP007766)
C. sp.	1	RM16704 (CP007769)
C. subantarcticus	2	LMG 24374 (CP007772), LMG 24377 (CP007773)
C. ureolyticus	1	RIGS 9880 (CP012195)

Table 3.2: Complete *Campylobacter* genome sequences analysed
All genomes were obtained from NCBI, the accession number for each genome sequence is given in brackets after the strain name.

can be viewed online at www.jackaidley.co.uk/phasome/campy, click on 'Gene Groupings' near the bottom of the page to view). On this figure, rows indicate homology groups whilst each column is a separate isolate. Green blocks indicate that the tract is located within the ORF, whilst orange blocks indicate that the tract is located close but not within the gene and grey blocks indicate non-PV homologues. The features labelled in red are: (1) example group from the *C. coli* core phasome not present in *C. fetus*; (2) example of a homology group shared between *C. coli* and *C. fetus*; (3) an example of a rare homology group, only present in one strain; (4) an example of a group that is present in all isolates but PV in only a few; (5) the indicated region contains the large, well-conserved core phasome of *C. fetus*; (6) note how there continue to be new homology groups identified even in the final isolate analysed.

3.5.1 Poly-G/C tracts are the most common form of putatively variable tract

There were 1944 poly-G/C tracts found in total (91.4% of the total number of SSRs), 139 poly-A/T (6.5%), while the remainder were all dinucleotide repeats. No longer repeat units were identified. Of the dinucleotide repeats discovered (44), 27 were AT or TA repeats, and 14 were CT or TC repeats. The longest dinucleotide repeat identified was a 13AG repeat.

77.3% of poly-G/C tracts were located within ORFs with just 4.6% of these being poly-C in the direction of coding. In contrast, just half (49.6%) of poly-A/T tracts were located within ORF, and, of these, 81.2% were poly-T in the direction of coding. 17/27 (63.0%) of AT and TA dinucleotide repeats were located with ORFs, and 11/14 (78.6%) of CT and TC repeats were located within ORFs. All other dinucleotide repeats (3) found were within ORFs.

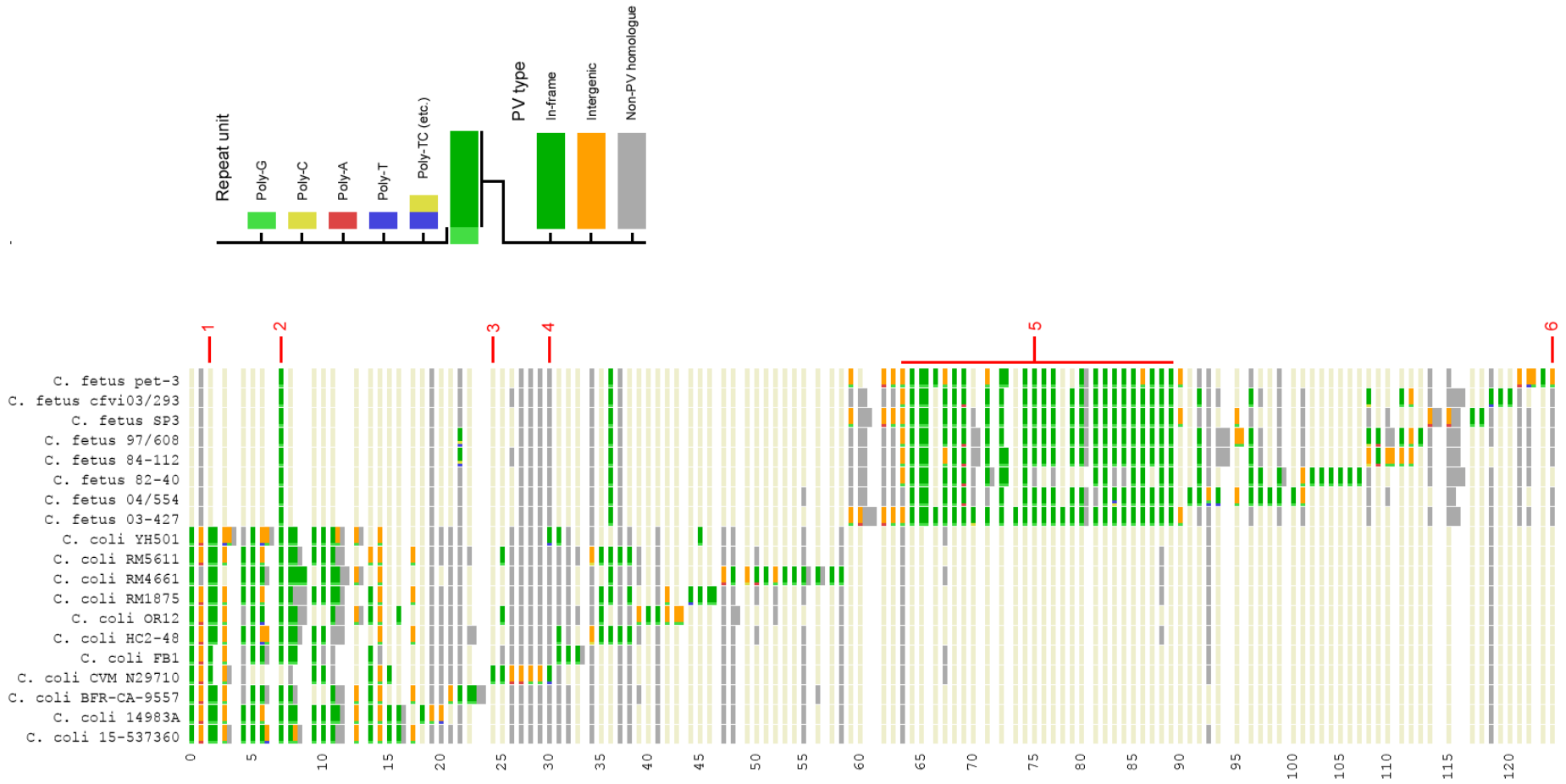

Figure 3.3: Example gene group graphic with *C. coli* and *C. fetus*
Figure illustrates homology of PV genes between isolates, each numbered row is a different homology group. Coloured blocks show presence or absence of PV genes: green is a tract located within the ORF, orange is a tract located nearby, and grey is non-PV homologue. Coloured bars at the start indicate the repeat unit on PV tracts. Faint beige bars are present to help readability and indicate absence of any homologue in that isolate. See text for meaning of the numbered points shown in red.

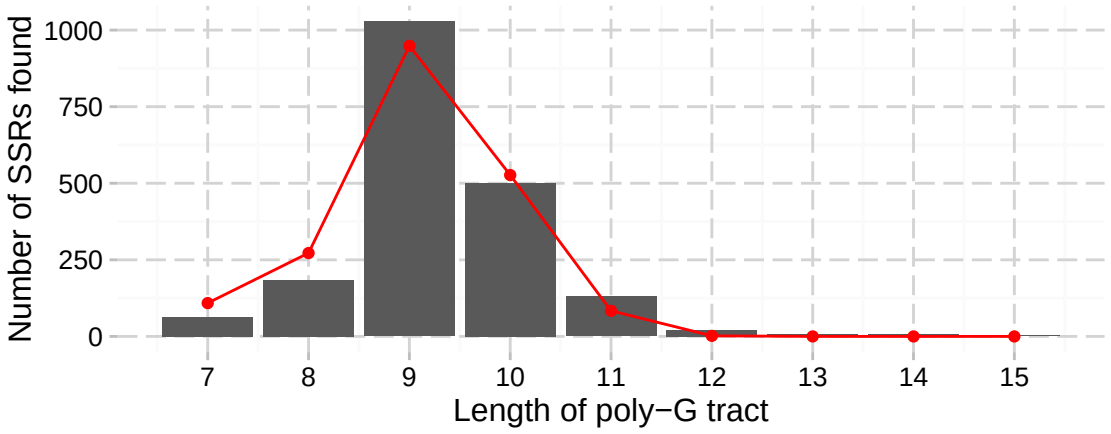

Figure 3.4: Distribution of poly-G tract lengths in *Campylobacter* genomes
Figure shows counts of poly-G tracts of each length found in the analysed *Campylobacter* genomes, from a total of 1944 tracts. Red line shows predicted distribution without selection based on empirical data of mutation rate for each tract length (see section 6.4.2). Poly-G tracts of length below 7 were excluded from the search. Tracts that are poly-C in the direction of coding are also included.

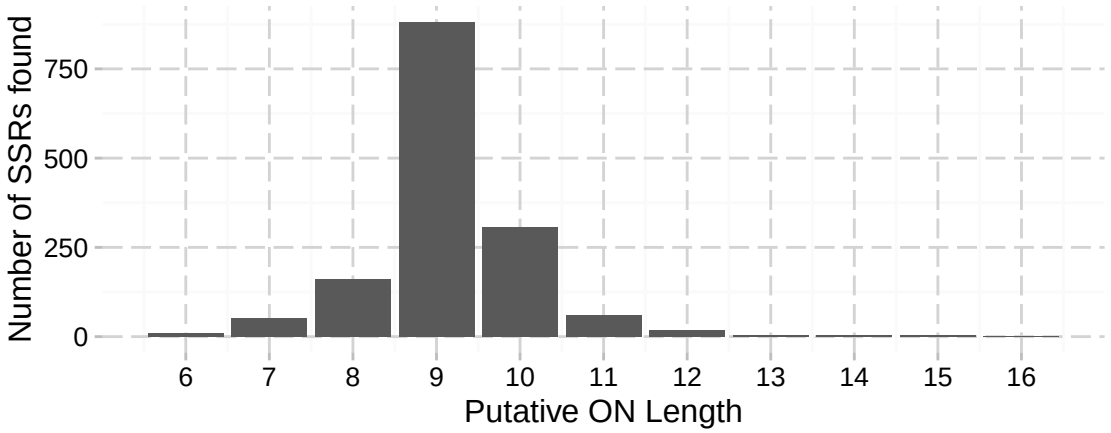

Figure 3.5: Distribution of putative ON lengths in *Campylobacter* genomes
Figure shows counts of nearest putative ON lengths found in the analysed *Campylobacter* genomes. This was determined by identifying the nearest tract length to the observed tract length that produces the longest continuous reading frame, and thus extends beyond the range of observed tracts in both directions. Only tracts located within reading frames are included (n=1503).

This supports the existing presumption that poly-G tracts are the primary source of phase variation in *Campylobacter* in general, however in two species — *C. hominis* and *C.ureolyticus* — this appears not to be the case. *C. hominis* contains just 2 poly-G tracts, 8 poly-A/T tracts of length 10 or longer and two dinucleotide tracts. Four of the poly-A/T are of length 10, on the cusp of the chosen threshold, and may or may not be genuinely variable. Three are of length 14-18, with two of these located in intergenic regions (one is potentially in a promoter region) and one located within the reading frame. The final tract is a 57T associated with a bacteriophage integrase and is likely to be bacteriophage associated rather than responsible for phase variation. *C. ureolyticus* contains two short poly-G/C tracts, two short dinucleotide tracts (AG7 and AT6) and a longer poly-T tract that is also phage associated. These two long poly-T tracts are the longest in the dataset, and the only poly-T tracts longer than 18 nucleotide. In both species only a single sequenced genome was included, however.

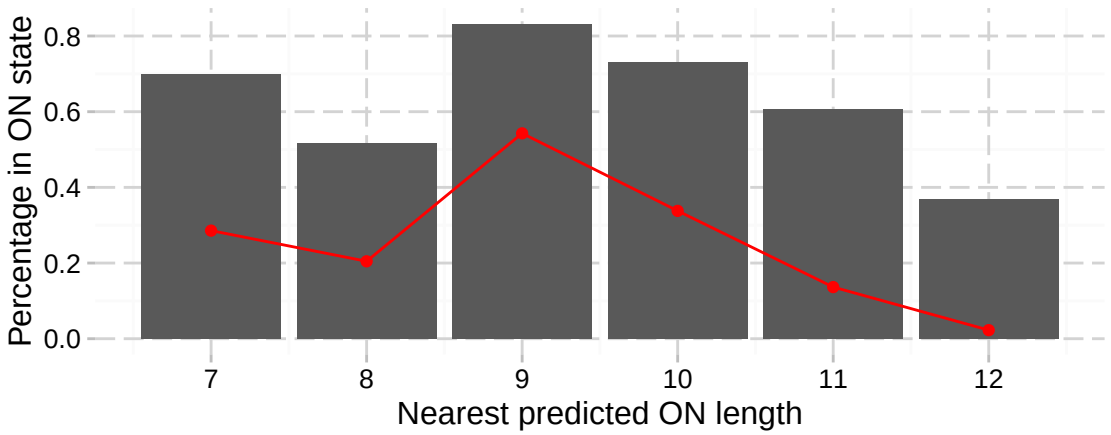

Figure 3.6: Proportion of tracts in the ON state by nearest predicted ON state
Data from collected *Campylobacter* genomes. Only tracts located within an ORF are included, and the predicted ON length was calculated as the tract nearest in length to the observed tract that leads to the longest protein product. Nearest predicted ON lengths with less than 10 representatives were excluded from the figure. The overlaid red line shows the predicted proportion that would be in the ON state if no selection was applied.

Across all genomes, the most common poly-G/C tract length of all strains was 9 (52.9% of Poly-G tracts), followed by 10 (25.7%), 8 (9.5%), 11 (6.7%), 7 (3.2%), and 12(1.1%). Other lengths (13-15) made up less than <1% of the population

(see figure 3.4). For those tracts where a putative ON-length could be deduced, these followed a similar profile, with a slightly greater bias towards 9 (figure 3.5). These ratios are close to those that would be expected by neutral flow based on models using empirical data of mutation rate as derived in section 6.4.2; these ratios are marked by the red line in figure figure 3.4. However, the proportion in the ON state for each length does not follow neutral patterns, as shown in figure 3.6 the proportion ON is higher than expected by neutral drift across all predicted ON lengths although it shows a similar pattern with a dip associated with the 8G length and a central peak around 9G. As the putative ON state is derived from the observed tract length, the proportion ON under neutral drift is calculated as the proportion of the putative ON state out of the three states including one above and one below the putative ON state. It is only with the 8G and 12G predicted ON lengths that the observed proportion ON is less than 60%.

There were between 5 and 81 PV genes identified in each strain, with distinct differences between species but also large variations within each species. In *C. jejuni*, the one representative of the *doylei* subspecies, 269.97 has over twice as many putatively variable SSRs as every other strain with 81 tracts. The distribution of the number of PV genes in each species is shown in figure 3.7. The number of homology groups follows a similar pattern (figure 3.8) as the ratio of PV genes to homology groups is similar in all species at between 1.0 and 1.5 except in *C. peloridis* in which it is 1.75.

3.5.2 Common functional groupings

Table 3.3 lists the twenty homology groups with the largest number of PV members identified by Phasome*It*, and a complete list of all homology groups is given in appendix B. The functions listed are automatically generated from annotation data and sometimes experimental knowledge can improve on these annotations.

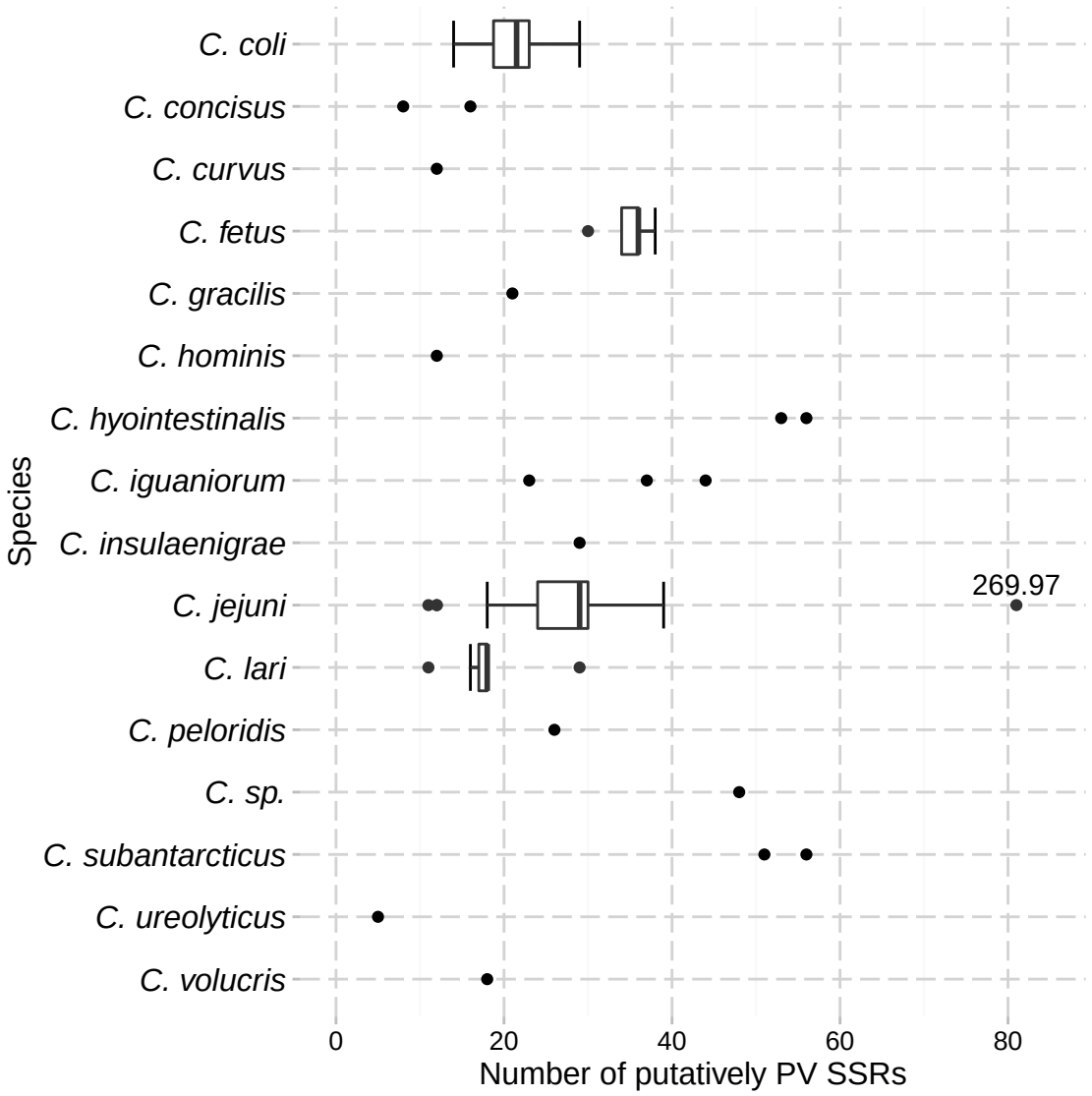

Figure 3.7: Number of PV genes per genome in each species
Figure shows the number of putatively PV genes identified for each species. For species with 5 or more isolates, box and whiskers are shown, for those with less each point is shown. These indicate median and interquartile range from 25% to 75%. The whiskers stretch to the further point within 1.5 interquartile ranges (IQRs) from the ends of the box, outlying points beyond this range are individually marked. The labelled point is for *C. jejuni* subsp. *doylei* strain 269.97.

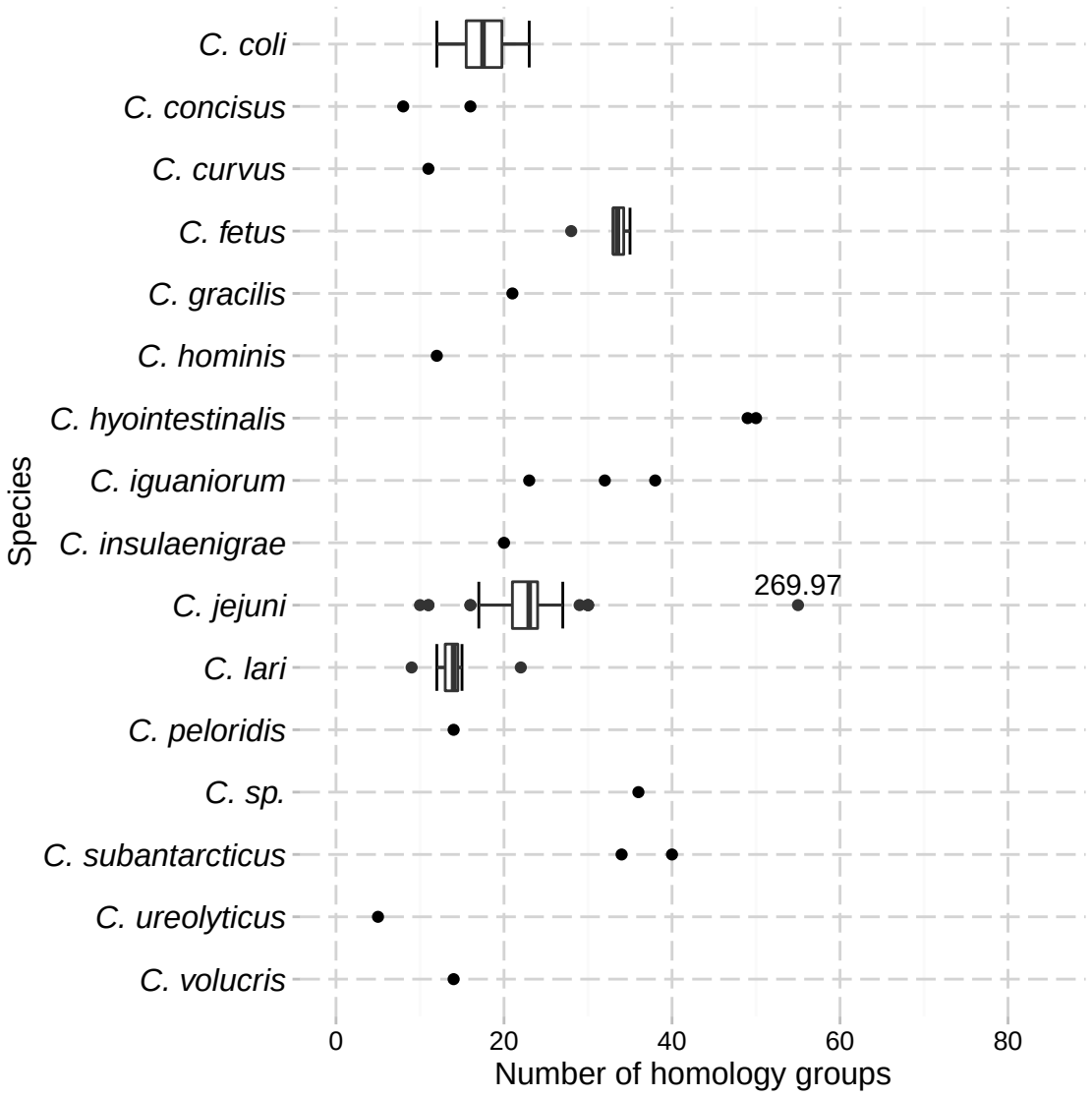

Figure 3.8: Number of homology groups per genome in each species
Figure shows the number of homology groups identified for each species. For species with 5 or more isolates, box and whiskers are shown, for those with less each point is shown. These indicate median and interquartile range from 25% to 75%. The whiskers stretch to the further point within 1.5 IQRs from the ends of the box, outlying points beyond this range are individually marked. The labelled point is for *C. jejuni* subsp. *doylei* strain 269.97.

In particular, the NCTC 11168 representative of the *cj1295* group is known to in fact be a flagellar modifying transferase (Hitchen *et al.*, 2010) and it is likely the rest of this group have similar function. Similarly, the homologues of the *maf1*, and *maf7* groups of genes in NCTC 11168 and 81-176 are known to modify the glycosylation of the flagellum (Karlyshev *et al.*, 2002, van Alphen *et al.*, 2008) and the *cj1421c* group are the group containing the MeOPN transferases that are the subject of much of this thesis. The *cj0170* group influences motility and thus may also modify the flagellum (Artymovich *et al.*, 2013). With this in mind it is apparent that the largest groups are mostly transferases of one sort or another involved in modifying the LOS, CPS, or flagellum. No information at all was gleaned from the annotation of two of the homology groups – *UPTC4110_0710* and *A911_t08342* – but the second of these is, in fact, a group of tracts located close to a tRNA.

Another common functional grouping is restriction modification systems, found in 9 homology groups. The most prevalent phase variable R/M homology groups are the *cj0031* group of type IIG restriction modification systems which has homologues in every *C. jejuni*, *C. coli* and *C. lari* strain except *C.jejuni* strain 32488, but is phase variable in just 12; and the *cj1051c* group which is not phase variable in the highly studied NCTC 11168 strain but has phase variable homologues in 21 other genomes from multiple species including *C. jejuni*, *C. coli*, *C. concisus*, and *C. fetus* as well as another 26 non-PV homologues (including that in NCTC 11168). There is a type III restriction modification system which has a PV mod subunit found in some *C. fetus* and *C. hyointestinalis* strains; and a separate group of PV mod subunits found in the two *C. subantarcticus* genomes with non-PV homologues in some *C. jejuni* and *C. coli* strains. There is also a group of phase variable S subunits from type I R/M systems with three representatives, with one found in each of *C. lari*, *C. hyointestinalis*, and *C. iguaniorum*; and a type II system found in two *C. iguaniorum* strains and one *C. concisus* strain. A further four PV R/M systems occur only in one strain.

Group name	#In frame	#Total PV	#Total	Putative function
maf7	200	203	233	carbonic anyhydrase[1]
cj1295	64	65	72	hypothetical protein (DUF2172 domain), putative M28 family zinc peptidase[1]
maf1	63	65	123	motility accessory factor[1]
cj0170	47	52	54	SAM-dependent methyltransferase[1]
cj1421c	45	52	55	putative sugar transferase[2]
ubiE_3	48	48	48	SAM-dependent methyltransferase
cj0045c	43	43	45	Hemerythrin-like iron-binding protein
cipA	36	36	39	Invasion protein CipA[2]
A911_07000	34	34	50	sugar transferase
UPTC4110_0710	30	31	45	No annotation data
cj1296	29	30	92	aminoglycoside N3'-acetyltransferase[1]
hxuB_1	0	25	36	Heme/hemopexin transporter protein HuxB precursor
ansA	0	24	78	L-asparaginase
cj0628	19	23	35	putative lipoprotein[2]
cjeI	21	21	47	restriction endonuclease
lgrA	21	21	29	formyl transferase domain protein
A911_t08342	0	21	75	No annotation data
CFT03427_1115	19	20	21	autotransporter domain protein
PJ18_06805	18	20	50	N-acetyl sugar amidotransferase
epsM	0	19	23	putative transferase

[1] Likely flagella modifying protein [2] Likely capsular modifying protein

Table 3.3: The twenty most common homology groups
The twenty most common homology groups ordered by the total number found associated with a putatively PV tract. The first number (#In frame) gives the number of tracts found with the SSR in the reading frame of the ORF; the second number (#Total PV)gives the total number of PV homologues found, including those where the SSR is close to, but not contained in the tract; and the third number (#Total) gives the total size of the homology group including non-PV tracts. Group name and function are automatically generated by Phasome*It*, but note that in many cases specific knowledge of function can be derived from the homologues in well studied species (see text). Likely functions indicated in footnotes are based on knowledge of the homologues in well studied strains, other transferases may also alter surface structures.

3.5.3 Two thirds of homology groups occur in just one isolate

536 homology groups were identified in total, with two thirds of these (347, 64.7%) occurring in just one isolate, and a further 163 in less than 5%. Just 4 were included in greater than 50% of the isolates. However, these isolates are collected from a range of species, with different levels of representation of each species so the data was re-analysed to look for homology groups associated with particular species.

3.5.4 Evidence of per species core phasome

The grouping of isolates by species can be seen by drawing a tree based on the presence or absence of homology groups. This is shown in figure 3.9. Note how the *C. jejuni* (teal and pink) isolates cluster to the left, the *C. coli* (black) isolates cluster to the top/right, *C. lari* (purple and pink) to the bottom and *C. fetus* (purple) to the lower right. The clustering of *C. jejuni* is disrupted by the clustering of the two *C. subantarcticus* isolates, the *C. volucris* isolate, and the *C. insulaenigrae* isolate with the NCTC 11351 and 269.97 (subspecies *doylei*) isolates of *C. jejuni*. This disrupted clustering is particularly striking as NCTC 11351 is the type strain of *C. jejuni*.

Four species (*C.jejuni*, *C. coli*, *C. fetus* and *C. lari*) are represented by 7 or more genomes and the following discussion centres on these species (see table 3.4 for a full list). *C. lari* shows the least conservation with just two homology groups in its core phasome and only one homology group (UPTC4110_0710 group) present in all *C. lari* strains but nothing is known about its function. This homology group is also present in the lari-like strains *C. peloridis*, *C. subantarcticus*, and *C. volucris*. It also has non-PV homologues in some *C. jejuni* strains and is present in multiple copies in most cases.

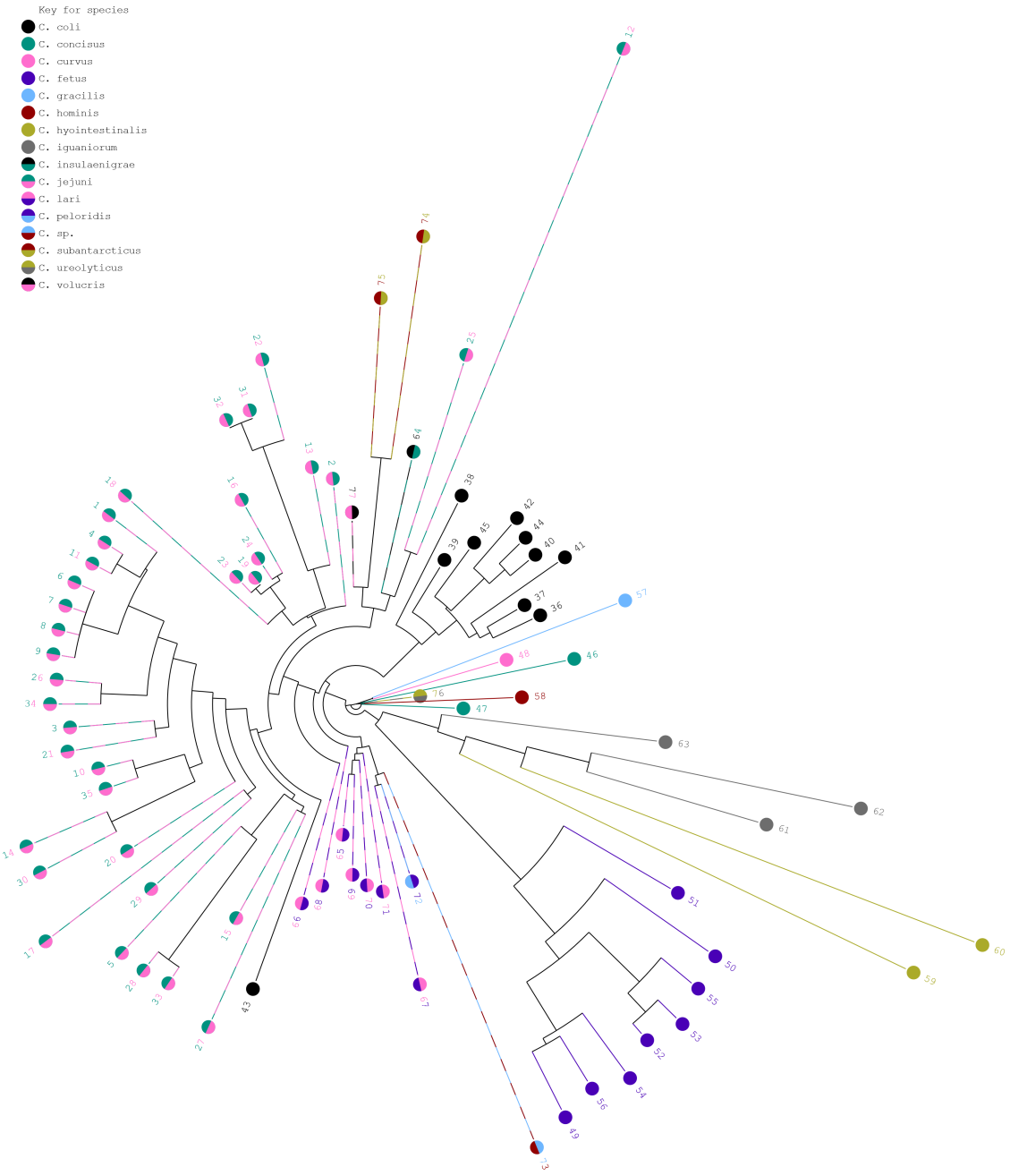

Figure 3.9: Neighbour joining tree of phasome similarity for *Campylobacter* species
Figure shows a neighbour joining tree of phasome similarity created by applying a Manhatten distance metric to binary lists of presence and absence of homology groups. Tree is coloured by species. Numbers indicate strains, a full list is given in appendix C, selected strains are: 1. *C. jejuni* NCTC 11168, 2. *C. jejuni* 81-176, 12. *C. jejuni* subsp. *doylei* 269.97, 25. *C. jejuni* NCTC 11351.

In % of strains	Group name	Function
		C. jejuni (n=35)
100.0	maf7	carbonic anyhydrase[1]
	Cj1295	hypothetical protein (DUF2172 domain), putative M28 family zinc peptidase[1]
94.3	Cj0045c	hemerythrin-like iron-binding protein
88.6	cipA	invasion protein CipA[2]
82.9	ubiE_3	SAM-dependent methyltransferase
71.4	hxuB_1	heme/hemopexin transporter protein HuxB precursor
68.6	Cj1421c	putative sugar transferase[2]
65.7	maf1	motility accessory factor[1]
	ansA	L-asparaginase
60.0	Cj0170	SAM-dependent methyltransferase[1]
		C. coli (n=10)
100.0	Cj0045c	hemerythrin-like iron-binding protein
	maf7	carbonic anyhydrase[1]
90.0	Cj1295	hypothetical protein (DUF2172 domain), putative M28 family zinc peptidase[1]
	murD	UDP-N-acetylmuramoylalanine–D-glutamate ligase
	N149_0842	hypothetical protein
	N149_0993	phosphoglycerol transferase
	hxuA	filamentous hemagglutinin
	vacA	autotransporter
70.0	Cj0170	SAM-dependent methyltransferase[1]
	maf1	motility accessory factor[1]
	lgrA	formyl transferase domain protein
	PJ17_06935	3-oxoacyl-ACP synthase
		C. fetus (n=8)
100.0	CFT03427_1684	hypothetical protein
	Cj1295	hypothetical protein (DUF2172 domain), putative M28 family zinc peptidase[1]
	ubiE_3	SAM-dependent methyltransferase
	CFT03427_0876	SAM-dependent methyltransferase
	CFT03427_0951	hypothetical protein
	CFT03427_1021	MCP-domain signal transduction protein
	CFT03427_1099	putative membrane protein
	CFT03427_1115	autotransporter domain protein
	CFT03427_1510	ATP-grasp domain protein
	CFT03427_1512	SAM-dependent methyltransferase
	CFT03427_1562	probable 3-demethylubiquinone-9 3-methyltransferase
	CFT03427_1573	hypothetical protein
	CFT03427_1574	hypothetical membrane protein
	CFT03427_1581	SAM-dependent methyltransferase
87.5	menA	1,4-dihydroxy-2-naphthoate octaprenyltransferase
	CFT03427_1442	transformation system protein
	CFT03427_1545	radical SAM superfamily enzyme, MoaA/NifB/PqqE/SkfB family
	CFT03427_1551	short-chain dehydrogenase/reductase family protein
	CFT03427_1554	methyltransferase
	CFT03427_1558	radical SAM superfamily enzyme, MoaA/NifB/PqqE/SkfB family (SPASM domain)
	CFT03427_1559	hypothetical protein
	CFT03427_1565	hypothetical protein
	CFT03427_1566	hypothetical protein
	CFT03427_1577	hypothetical membrane protein
75.0	CFT03427_1556	formyltransferase domain-containing protein
62.5	CFF04554_0871	putative type II secretion system protein
	CFF04554_1255	4HB_MCP sensor-containing MCP-domain signal transduction protein
		C. lari (n=7)
100.0	UPTC4110_0710	hypothetical protein
71.4	UPTC4110_1471	MCP-domain signal transduction protein

[1] Likely flagella modifying protein [2] Likely capsular modifying protein

Table 3.4: Core phasome of *C. jejuni*, *C. coli*, *C. fetus*, and *C. lari*
Table shows homology groups present in 60% or more of the genomes from each species. The percentage of genomes in each species containing the homology group is shown in the first column. Each group is assigned a name based on the first gene name found in the PV genes in the group, or failing that the first locus name found. These names are preferentially chosen from a manually curated order that favours well-studied species. The functional assignment is based on annotation data associated with the genomes and is automatically obtained as described in the main text.

At the other end of the spectrum, *C. fetus* has 27 genes in its core phasome, with 14 present in all 8 genomes. Since strains of *C. fetus* contained 28-35 (mean 32.9) homology groups, this means that a majority of homology groups present in each strain are core phasome, this is despite the genomes included being drawn from three subspecies: *fetus* (n=2), *testudinum* (n=3), and *venerealis* (n=3). *C. jejuni* and *C. coli* have 12 and 10 homology groups in their core phasome, respectively.

There is some overlap between these core phasomes; see figure 3.10 which indicates the overlap between four species. Five homology groups are shared between the core phasome of *C. jejuni* and *C. coli*: the *cj0045c* group, the *cj0170* group, the *cj0617* group, the *cj1295* group (which is also found in the core phasome of *C. fetus*, *C. hyointestinalis* , *C. iguaniorum*, and *C. subantarcticus*), and the *maf1* group.

There are four other species represented by 2 or 3 isolates – *C. concisus*, *C. hyointestinalis*, *C. iguaniorum*, and *C. subantarcticus* – which is too few to give a clear indication of the core phasome of these species, however some general observations can be made. Of these, *C. concisus* seems to have no homology groups preserved across the species from the two genomes included. *C. subantarticus* (n=2) and *iquanorium* (n=3) both have putative core phasomes of about 20 groups, with *C. subantarticus* having a number of groups in common with *C. coli* and *C. jejuni*. *C. hyointestinalis* (n=2) has a smaller core phasome of 13 groups, most of which are unique to *C. hyointestinalis* but also contains the *cj1295* group.

Common themes emerge among these core phasomes. In each of the four well-represented species the core phasome contains at least one homology group which not only occurs in all isolates but also has multiple copies in each isolate. Interestingly, despite their overall prevalence, no identified restriction-modification system is among the core phasome of these species, whereas potentially surface-modifying transferases are frequently represented among the core phasome and flagellum affecting groups are particularly widespread.

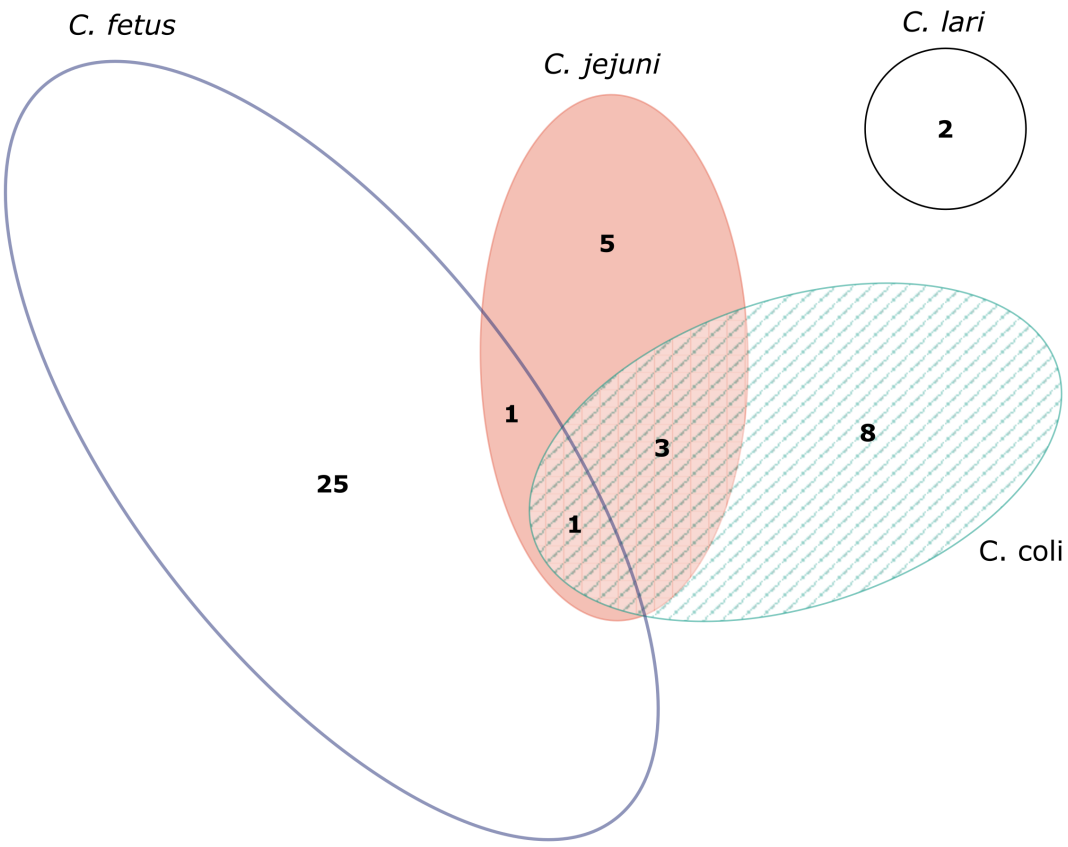

Figure 3.10: Overlap in core phasome across four species
Euler diagram shows overlap in core phasome between four species: *C. coli*, *C. fetus*, *C. jejuni*, and *C. lari*. Overlapping areas show shared homology groups in the core phasome, and the area of each region is proportional to the size of the set represented.

3.5.5 Many PV genes have non-PV homologues

55.6% (298) of homology groups contain not just PV genes but also non-PV homologues. In some cases, there are both PV and non-PV orthologues in the same genome. These non-PV homologues can have very high levels of sequence similarity (> 90%) but critically differ at the site of the SSR. In some cases, the non-PV homologue contains a shorter, or interrupted, poly-G tract however in other cases there is no sequence that resembles a variable tract. Examples of this are illustrated in figure 3.11.

Shorter tract (*clpX* group):

```
              I  I  E  G  S  L  V  N  I  P  P  R  G  G  R  K  H  P  N  Q  E  F
NCTC 11168    ATTATCGAAGGAAGTTTGGTAAATATTCCACCAAGGGGGGGGAGAAAACATCCAAATCAAGAGTTT
              ||||||||||||||||||||||||||||||||..||||||.|||||||||||||||||||||||
81-176        ATTATCGAAGGAAGTTTGGTAAATATTCCACCAAAAGGGGGAAGAAAACATCCAAATCAAGAGTTT
              I  I  E  G  S  L  V  N  I  P  P  K  G  G  R  K  H  P  N  Q  E  F
```

Disrupted tract (*CJJ81176_0758* group):

```
              L  I  T  A  A  F  T  D  R  G  G  G  F  N  N  Q  T  I  L  K  E  N
4031          TTAATTACTGCGGCGTTTACTGACAGGGGGGGGGGGATTTAACAATCAAACCATTTTAAAAGAAAAC
              ||||||||||||||||||||||||||||||.||||||||||||||||||||||||||||||||||
NCTC 11168    TTAATTACTGCGGCGTTTACTGACAGGGGGGAGGGATTTAACAATCAAACCATTTTAAAAGAAAAC
              L  I  T  A  A  F  T  D  R  G  E  G  F  N  N  Q  T  I  L  K  E  N
```

Dissimilar sequence (*kfoC* group):

```
              G  L  L  R  A  R  Y  E  G  V  K  A  A  G  G  G  Y  I  M  F  L  D
NCTC 11351    GGTCTTTTAAGAGCTAGATATGAAGGAGTTAAGGCAGCTGGGGGGGGGATATATTATGTTTTTAGAC
              ||||||||||||||.||||||||||||.||.||.|.|||.        .||||||.||||||||||.
81-176        GGTCTTTTAAGAGCAAGATATGAAGGTGTGAAAGTAGCAAACTCTCCTTATATAATGTTTTTAGAT
              G  L  L  R  A  R  Y  E  G  V  K  V  A  N  S  P  Y  I  M  F  L  D
```

Figure 3.11: Comparison of region around PV tract against non-PV homologues
Figure shows examples of three possible cases of how PV and non-PV homologues differ. In each case, the PV homologue is shown on the top with the translated sequence above, and the non-PV homologue is shown below with the translated sequence below. Amino acid sequence differences are highlighted in red. All three are drawn from strains of *C. jejuni*.

These homology groups can be present in all strains but only phase variable in one. An example of this is the homologue of *cj1120c* in the *C. jejuni* subsp. *doylei* isolate, 296.97 which has an 11T tract located near to the start of the gene which

is absent in other isolates but this gene is present in every *Campylobacter* genome and variously annotated as a UDP-N-acetylglucosamine 4,6-dehydratase and a N-acetyl-hexosamine dehydratase, among other similar functional annotations.

3.5.6 Putatively variable SSRs in plasmids

Plasmids are present in a number of these isolates but were excluded from the main analysis. Separate analysis of the plasmids alone reveals the presence of some putatively variable SSRs that form six homology groups.

The pVir plasmid found in the 81-167, IA3902 and 00-0949 strains of *C. jejuni*, and the closely related pCj2 plasmid found in 01-1512, contains a single poly-C tract located within the gene, *cjp27* and a long poly-A tract located upstream of *cjp28*. The gene *cjp27* has some homology to the replicative protein RepE (Bacon *et al.*, 2002), and is positioned between *cjp26* and *cjp28* which are homologous to other replicative proteins. However, the poly-C tract is located close to the end of the protein and results in a difference of at most 4 amino acids in length so it is not clear whether this produces a phase variable effect. The tract is of different length in the different sequenced strains (10C in IA3902, 11C in the others) suggesting that it does undergo the expected length variability. The location of the poly-A tract does not suggest any obvious functional role and its length varies between 15A and 21A in the sequenced plasmids.

3.6 Phasome analysis with isolates of known host attribution

An important question about the phasome is whether the observed variation, or any particular homology group, is associated with particular aspects of the

organism's environment. To investigate this, Phasome*It* was applied to a second dataset consisting of incomplete genome sequences of *C. jejuni* and *C. coli* isolates of known source. This collection was obtained between 2001 and 2009, with isolates collected from around the world but mostly in the UK. Each isolate comes from a known source – i.e. chicken, cattle, human, etc. – and thus represent isolates from throughout the food chain from farm to disease isolates but also some environmental isolates obtained from wild birds, farmland, and riparian environments. To this set the well annotated sequence of NCTC 11168 (Parkhill *et al.*, 2000, Gundogdu *et al.*, 2007) was added to provide a known comparison for *C. jejuni*, and well annotated strain 15-537360 of *C. coli* (Pearson *et al.*, 2013) which additionally contains all genes previously identified in the previous section as being in the core phasome of *C. coli*.

3.6.1 Similarities and contrasts with the complete genome set

The analysis of this new dataset re-iterates many of the points of the previous analysis. Poly-G tracts make up the largest proportion of the SSRs identified (88.4%, 3193 out of 3611 in total), and most of these are located in ORFs (81.8%), similar to to the complete genome analysis. There was also a single pentanucleotide repeat identified, a 5ATATT repeat, and nine longer (11-14 repeat) CT repeats, but these were very rare compared to the other repeat types. The lengths of the poly-G tracts were also similar in distribution, 52.2% were length 9, 29.5% length 10, and 7.0% length 8. The distribution was slightly wider, with lengths up to 20G observed. As before the nearest predicted ON lengths were similar to the distribution of observed tract lengths, however, the proportion of tracts that were in the ON state is lower than in the complete genome set, and closer to the proportion predicted by neutral drift (figure 3.12).

The number of PV tracts per genome was similar to that observed in the complete

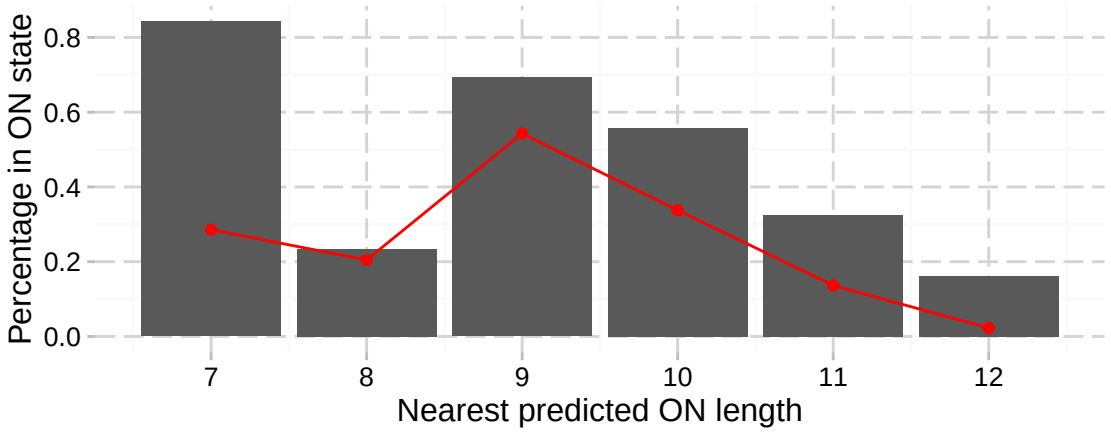

Figure 3.12: Proportion of tract in the ON state by nearest predicted ON state in host-associated genomes
Data from collected Campylobacter genomes. Only tracts located within an ORF are included, and the predicted ON length was calculated as the tract nearest in length to the observed tract that leads to the longest protein product. Nearest predicted ON lengths with less than 10 representatives were excluded from the figure. The overlaid red line shows the predicted proportion that would be in the ON state if no selection was applied.

genome set, with a median of 20 (16-26 interquartile range) for *C. jejuni* and a median of 15 (11-17 interquartile range) for *C. coli*, as was the number of homology groups that these PV genes fell into. The sequencing and annotation pathway of this collection also resulted in the identification of a number of SSRs containing ORFs less than 20 amino acids long, these are too short to effectively search for homologues of by BLAST and were excluded from the analysis as they are believed to be artefacts of the sequencing and annotation protocol. There were 304 of these ORFs in total, at an average of 1.61 per genome.

3.6.2 Isolates group by ST complex based on homology groups

As before, trees were constructed based on the presence or absence of homology groups. These trees showed clear separation by species (data not shown) and a strong separation by ST-complex (figure 3.13). There is a small group of genomes (marked on figure 3.13) that group together and contain a mix of species and ST-complexes. This group represents poorly sequenced isolates rather than having any biological significance. The genes used to determine ST-complex were checked to

ensure that they themselves are not phase variable (data obtained from pubMLST – Jolley and Maiden 2010) and thus this result is not simply a result of the same genes being used.

3.6.3 There is weak association between the phasome and host type

Next, the tree was annotated with the source of each isolate (e.g. cattle, chicken, human, etc.) which revealed some local clustering by type but each source attribution was found widely distributed across the tree (data not shown). Because some ST-complexes are only associated with particular hosts the observed local clustering may be a result of the clustering by ST-complex, and any more significant association could be masked by the clustering by ST-complex. Accordingly, the data was re-analysed among each ST-complex. Representation by ST-complex is inconsistent with some represented by larger numbers than others, but three ST-complexes were represented in larger numbers: the ST-21 complex (n=28), the ST-45 complex (n=28), and the ST-828 complex (n=41). ST-21 complex and ST-45 complex are ST-complexes of *C. jejuni*, while ST-828 is a complex of *C. coli*. All three have been isolated from chicken, cattle and human disease isolates. Homology group trees are shown in figures 3.14, 3.15, and 3.16. Because *Campylobacter* is a zoonotic disease transmitted from farmyard to human hosts rather than passed directly between human hosts it would be expected that disease isolates would cluster within other host attributions and this can be observed on all three figures. Cattle isolates are rare in the ST-21 complex and do not show signs of clustering together. There is some visual clustering in ST-45 and ST-828 by host type, however this is not statistically significant in ST-45 ($p > 0.5$, tree-based scan statistic with conditional Poisson model). Only in ST-828 do the cattle host isolates cluster ($p < 0.05$, tree-based scan statistic with conditional Poisson model).

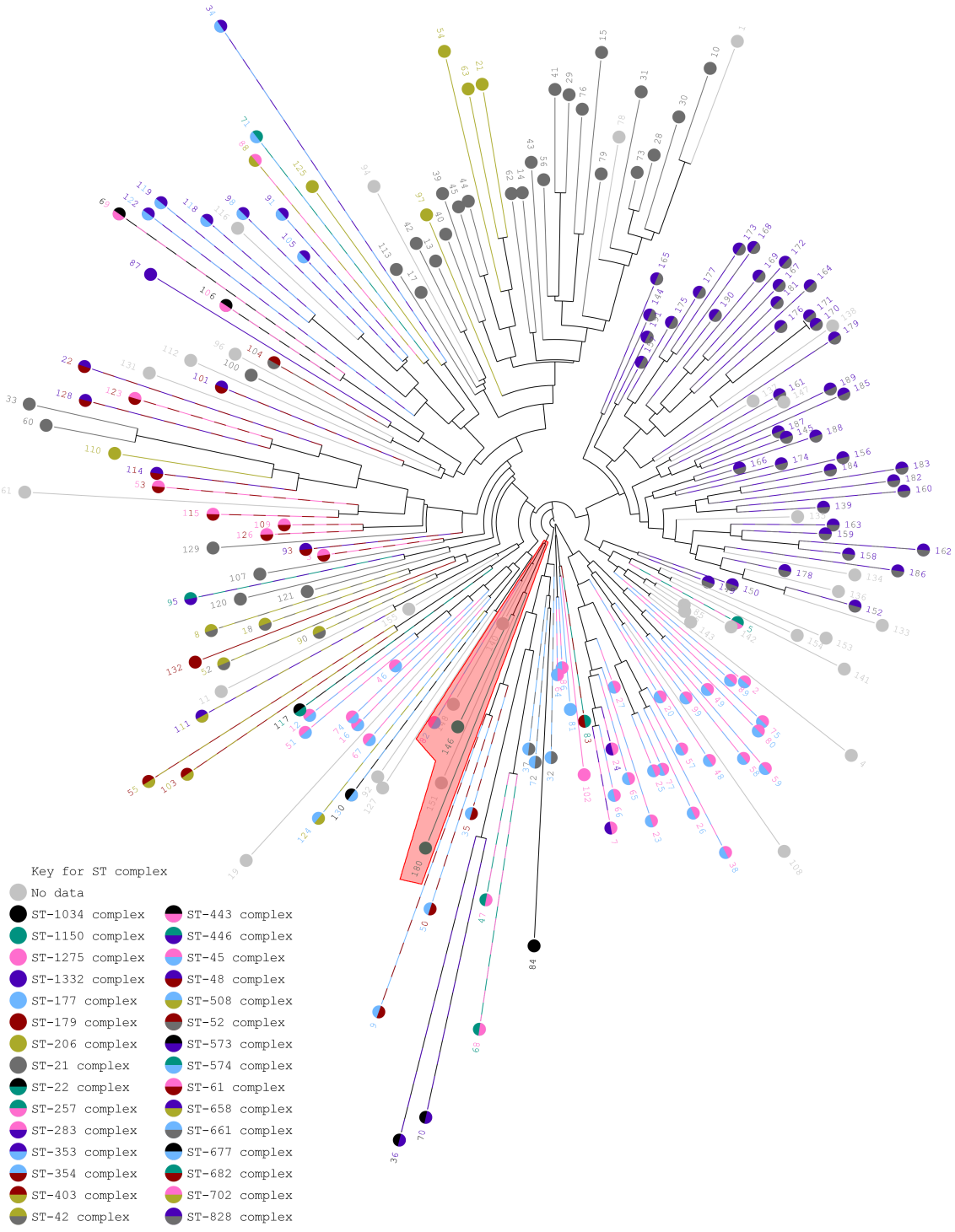

Figure 3.13: Tree based on homology groups shows clear separation by ST-complex
Figure shows a neighbour-joining tree using a Manhattan distance between isolates based on presence or absence of homology groups. Tree is coloured to indicate ST-Complex of isolates. A group of poorly sequenced genomes is indicated in red. Numbers are isolate numbers and carry no biological significance.

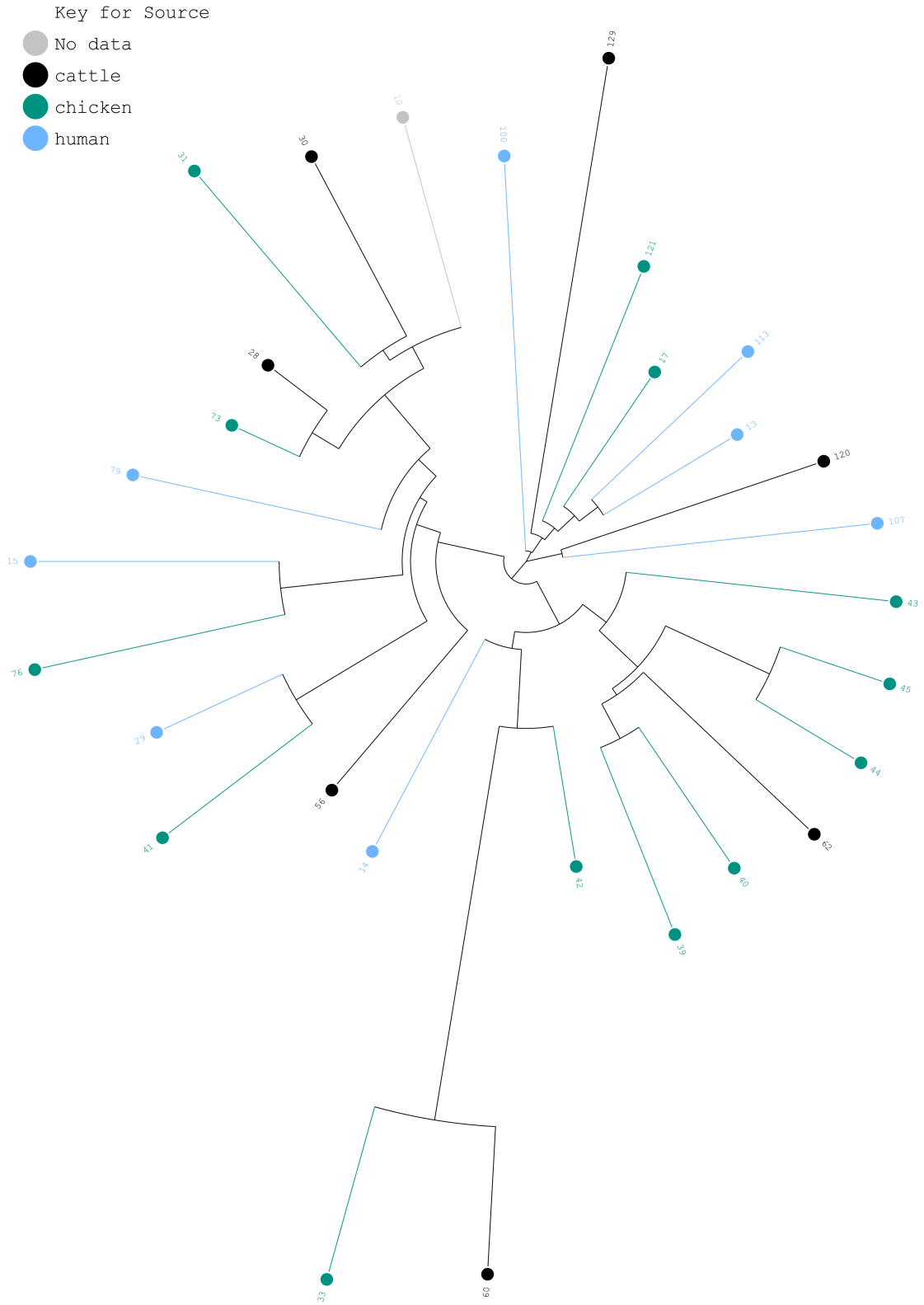

Figure 3.14: Host association in the *C. jejuni* ST-21 complex
Figure shows a neighbour-joining tree using a Manhattan distance between isolates based on presence or absence of homology groups in *C. jejuni* isolates from the ST-21 complex. Tree is coloured according to the source the isolate was originally isolated from. Any apparent grouping is not statistically significant (tree-based scan statistic with conditional Poisson model). Numbers are isolate numbers and carry no biological significance.

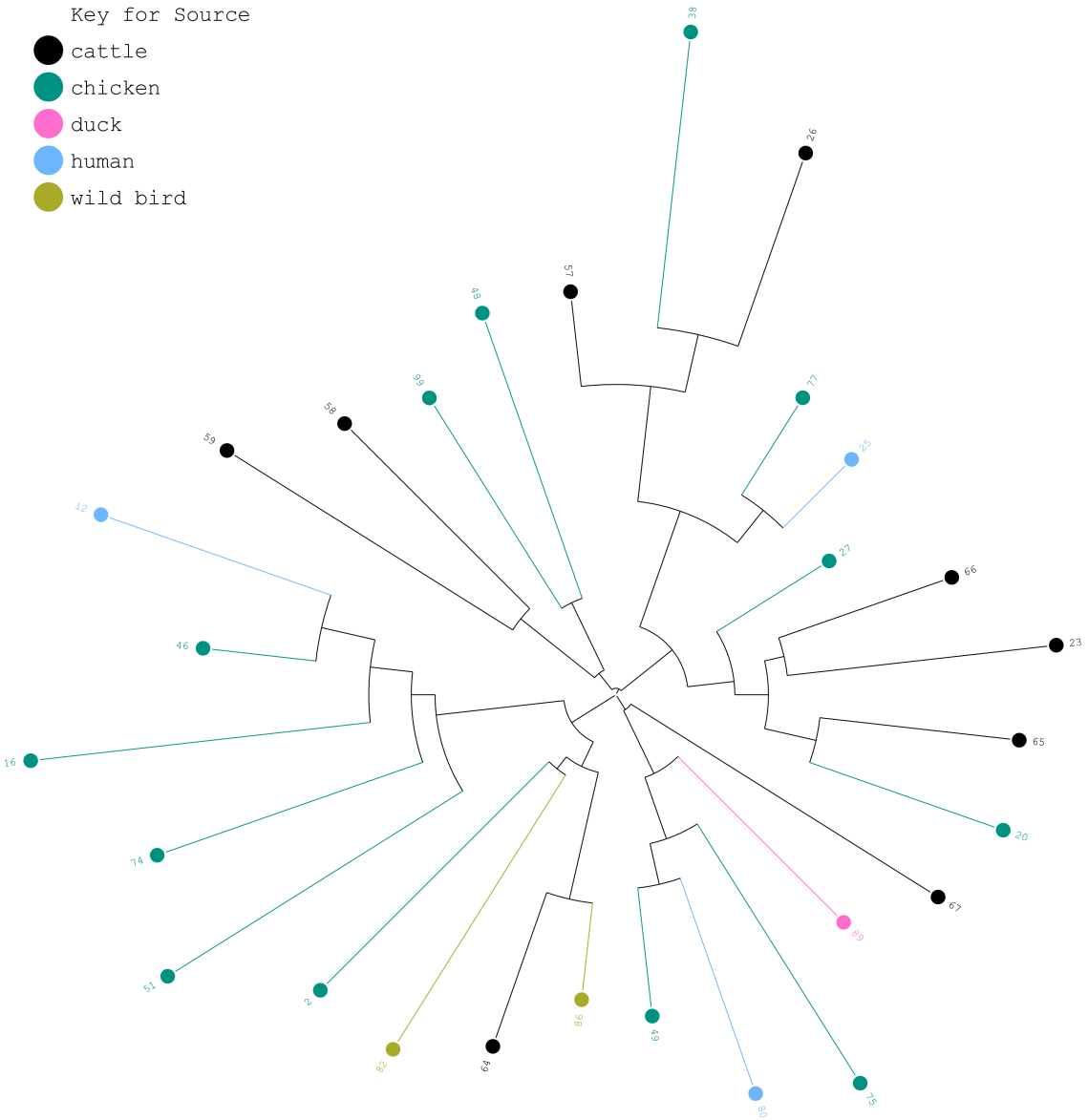

Figure 3.15: Host association in the *C. jejuni* ST-45 complex
Figure shows a neighbour-joining tree using a Manhattan distance between isolates based on presence or absence of homology groups in *C. jejuni* isolates from the ST-45 complex. Tree is coloured according to the host the isolates were obtained from. Any apparent grouping is not statistically significant (tree-based scan statistic with conditional Poisson model). Numbers are isolate numbers and carry no biological significance.

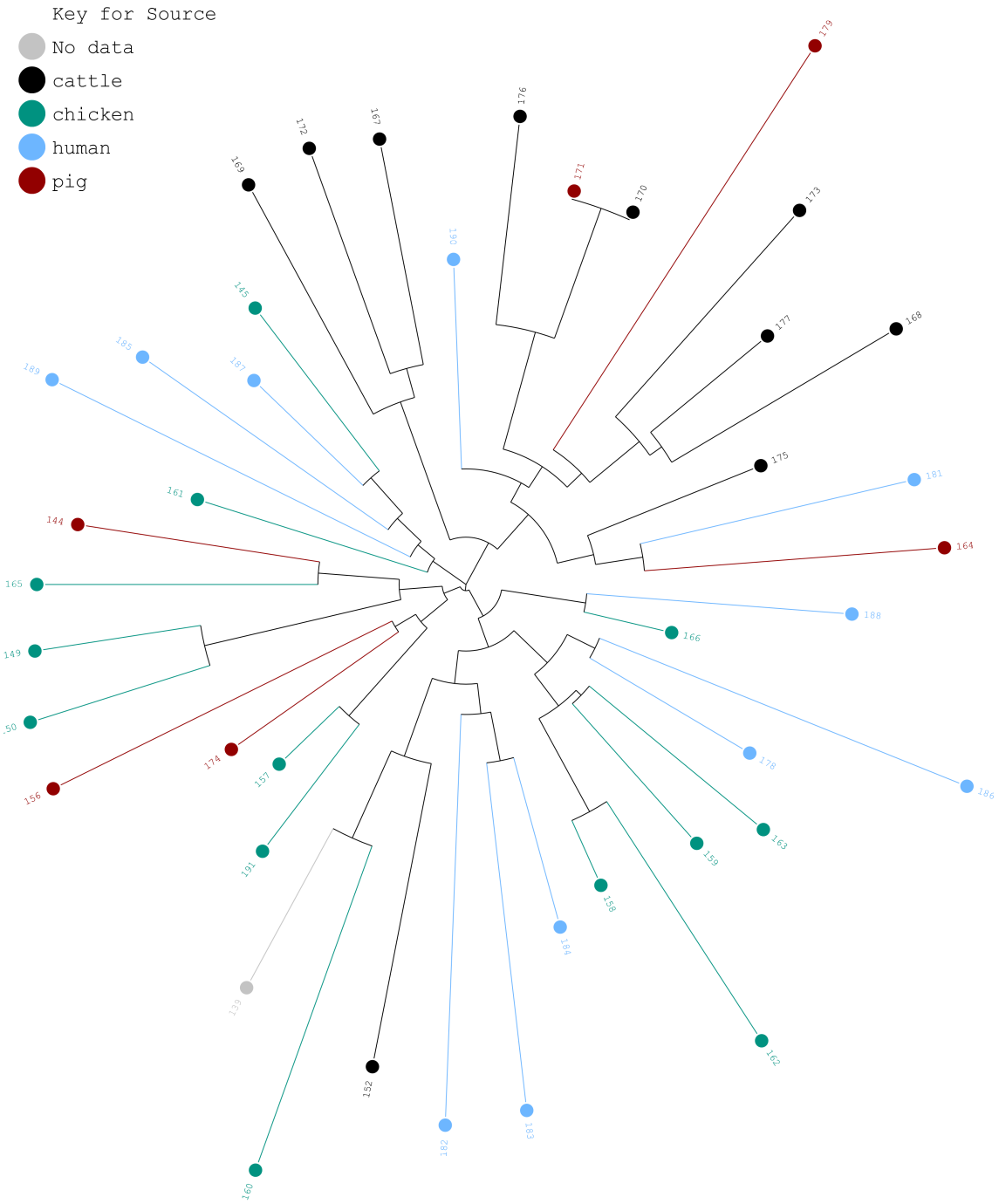

Figure 3.16: Host association in the *C. coli* ST-828 complex
Figure shows a neighbour-joining tree using a Manhattan distance between isolates based on presence or absence of homology groups in *C. coli* isolates from ST-828 complex. Tree is coloured according to the host the isolates were obtained from. Isolates from cattle show significant grouping within the tree ($p < 0.05$, tree-based scan statistic with conditional Poisson model). Numbers are isolate numbers and carry no biological significance.

Chapter 4

Phage and Sera have opposite selective effects on phase variable expression of *cj1421c*

Note: The program PSAnalyse, in section 4.3, has previously been published in Lango-Scholey *et al*. (2016).

4.1 Summary

In order to move towards the eventual aim of developing a cyclical selection assay it was necessary to first confirm that selection could be reliably achieved in both directions for the identified target gene, *cj1421c* and that this process was not producing selection for any other genes. The chosen selective agents were phage F336, against which a change from *cj1421c*-ON to *cj1421c*-OFF has been shown

to produce resistance, and human sera against which the MeO*P*N group has been shown to provide protection in other *C. jejuni* strains. These experiments were initially carried out using wild type strains and, as expected, confounding selection for the similar phase variable MeO*P*N transferase *cj1422c* was observed with both selection using sera and counter-selection using phage F336. However, in both assays selection was seen only for these two genes while the appropriate selection for *cj1421c* was also observed. A knockout strain, Δ*cj1422c*::RDH315, was constructed without *cj1422c* and used to confirm phage selection for *cj1421c*-OFF operates in this strain. Evidence of differences in selection strength between samples of pooled human serum were also identified. Additionally this chapter describes a computer script, PSAnalyse, that was developed in order to facilitate the rapid analysis of the large datasets produced by these experiments.

4.2 Introduction

The previous chapter assessed the wide diversity of phase variable tracts present across isolates of *Camplobacter* and the variation in the unit sizes and repeat numbers of SSR tracts both between isolates and between different phase variable loci in the same isolate. Differences in SSR repeat number result in different frequencies of phase variation in *C. jejuni* (Bayliss *et al.*, 2012) and other species (De Bolle *et al.*, 2000, Morel *et al.*, 1998). While there is theoretical data suggesting that these differences may be biologically relevant (Palmer *et al.*, 2013, and references) there has been no published experimental investigation into the functional role of these differences in frequency resulting from tract length although experiments in similar systems have been performed (Acar *et al.*, 2008). The rest of this thesis will concentrate on the development of an experimental model with which to study the biological significance of these differences alongside a parallel *in silico* model against which experimental findings can be directly compared.

The aim is to develop an experimental system where an alternating selective pressure can be applied to a specific phase variable locus. For this to be achieved two forms of selection acting in opposite directions on a single gene needed to be identified, and these forms of selection need to act only on the target gene. The chosen target gene is *cj1421c*, a phosphoramidate transferase (see section 1.7.1). Multiple possible selective agents on this gene have been described in the literature.

Sørensen *et al.* (2011) showed that *in vitro* growth of sensitive NCTC 11168 with the F336 phage allows resistant colonies to be isolated. They further showed that this resistance was produced by a change in the the phase variable state of *cj1421c* from the ON to the OFF state. Sørensen *et al.* (2012) reported that the resistance of the bacterium to phage F336 depends not just on *cj1421c* but also on the second MeO*P*N tranfserase *cj1422c* with the ON state of *cj1422c* conferring resistance regardless of the state of *cj1421c*. They further showed that a third gene *cj1426c* has a mild protective effect reducing the EOP by approximately 50%. Experimentally, they went on to show that presence of phage F336 during *in vivo* passage selected for a change in *cj1421c* expression state from a functional (9G) state to a non-functional (10G) state and that it was possible to produce resistant mutants at high frequency *in vitro*. Thus it is expected that *in vitro* addition of phage F336 during growth of NCTC 11168 will select for resistant states of the two phase variable phosphoramidate transferases.

Selection in the opposite direction is also possible. Work from Champion *et al.* (2010) suggested that a phosphoramidate group was required for invasion of *G. mellonella* larvae. However later work, from van Alphen *et al.* (2014), showed that although knocking out the phosphoramidate synthesis pathway limited invasion of *G. mellonella* larvae, there was no effect from the deletion of the transferases. These results suggest that there is another aspect of the MeO*P*N biosynthesis pathway that facilitates invasion in this model. However, van Alphen *et al.* (2014) also investigated a range of other biological impacts of the phosphoramidate group and

showed that the residues have a protective effect against human sera and increase the rate of piglet colonization. Due to the relative ease of applying serum-based selection in the laboratory setting, serum selection was chosen as the selective agent.

Since with both phage and sera selection it is expected that the gene *cj1422c* will have a confounding effect on *cj1421c* selection, knockout strains will be needed for the final cyclical selection assay. However the first step was to characterise the strength and effectiveness of these selective agents in wild type strains and this investigation composes the majority of this chapter. Before moving on to the experimental data I first describe the development of a tool for rapid analysis of phase variation data: PSAnalyse.

4.3 Development of PSAnalyse facilitates high throughput analysis of phase variable tracts

In brief, the length of a poly-G tract can be determined by analysis of PCR products with fluorescent labels using capillary electrophoresis alongside a set of markers of known length labelled with a different fluorescent marker (see section 2.3 for full description). The data produced is then analysed using Peak Scanner™ which produces a visual trace of the electrophoresis and precise data on every peak identified. The desired peaks can be identified by dye type and product size followed by comparison of this size to the equivalent peak in samples of known tract length and calculation of the length of the tract in the new samples.

Quantification of poly-G track length in strain NCTC 11168 produces a very large quantity of data when applied to experimental datasets. With 28 tracts amplified,

90 samples on a plate produces over 1600 individual datapoints, while each experiment described later in this thesis typically utilized 3 to 7 plates, potentially leading to over 100,000 individual datapoints in total. In previous work these data had to be individually identified and quantified by hand using the process outlined above. After identification of each datum, it was then necessary to manually curate this data to produce ON/OFF scoring data. Due to the highly time consuming and potentially error-prone nature of this process, I developed the PSAnalyse script to automate data processing and to produce both human and machine readable outputs from the datasets. This script sharply reduces both the time required for analysis and the number of errors introduced.

4.3.1 Analysis process

PSAnalyse carries out a post-analysis of PCR products analysed by GeneScan™ and Peak Scanner™ v1.0 or v2.0. Peak Scanner™ takes a .fsa format file produced by a Applied Biosystems capillary electrophoresis instrument and produces a list of all fragments identified in the sample after applying an algorithmically determined cutoff on signal strength (peak height) to screen out background noise. The size of these peaks is calculated in base pairs by calibration to a size standard with labelled fragments of known size. This list can be exported as a tab-deliminated file containing detailed information for each peak which can be read by PSAnalyse.

In order to convert this into tract length data, PSAnalyse must first know where it expects to find the peaks. This is obtained by taking a single sample and both passing it through the GeneScan™ process and Sanger sequencing to determine the tract lengths. This creates a mapping between the observed fragment sizes and a known tract length which can be turned into a comma separated values (CSV) control data file for PSAnalyse. These control data files contain the known tracts along with the expected peak size; the tract length this size corresponds to; the

tract length that corresponds to the ON state for each tract; and potentially a step distance in base pairs between different ON states (e.g. the 3bp gap between 9G-ON and 12G-ON that produces the same reading frame). These files are referred to as peaksets within the script. PSAnalyse uses this information to collect all peaks within a specified range of the expected size of a target peak (default $\pm$3.5bp) and selects the highest peak as the matching peak. The script determines the size of the peak in bp and compares this to the expected peak position to determine the tract length. This length is then compared to the ON peak length to generate an ON/OFF score for the tract, represented by 1 for ON and 0 for OFF. Slippage during PCR amplification, as well as pre-existing variation within the colony, mean that there are normally smaller "side" peaks located 1bp above and below the main peak. PSAnalyse also identifies these side peaks and compares their area to the main peak. In some cases the main and side peaks are similar in size and thus a single tract length cannot be confidently assigned. To exclude these ambiguous cases a cutoff is applied to the ratio of the main peak area to the side peak areas (default 1.5) below which the peak is marked as uncertain.

In principle the difference between the peak positions in base pairs should always change by exact multiples of 1 as the tract length differs in size, however in practice there is a small error in every measurement. To account for this, the script takes a maximum acceptable error describing how far from the expected value the peak can be and still be scored (default $\pm$0.33bp). A substantial part of this error however is a batch effect in which a peak is shifted in all samples by approximately the same amount. To correct for this a known sample run on each plate can be used as a calibration control. The script is passed the known lengths of the tract for each expected peak and the name of the well containing the control. It uses this calibration data to calculate new expected positions for every peak and applies these instead of the passed values for every other well. This substantially reduces the error and allows many more peaks to be correctly called.

The list of peaks generated by Peak Scanner™ also contains a large number of spurious data points. The largest of these are generated during the initial phase of the electrophoresis and probably result from unbound primers and prematurely terminated PCR products. Since most of these spurious peaks occur at much lower sizes than the target peaks they are easily ignored since they do not fall close to the expected peak sizes, but a significant fraction are still observed. The noise reduction filters applied by Peak Scanner™ exclude most small spurious peaks but this filtration still lets a substantial number of small peaks through and, inevitably, some of these fall close to the expected peak sizes. Therefore PSAnalyse also includes a minimum peak size criteria to exclude peaks that could potentially be false positives.

To make the analysis of the output data easier, PSAnalyse will also accept a "names file" which contains human readable names for the samples in each well. These are then included in the output data. This file can be simply created in Excel and saved in 'CSV (comma delimited) (.csv)' format with the name in each well corresponding to the similarly named cell in Excel (i.e the name in cell B11 is applied to the sample in well B11). Unfortunately, however, Excel uses a left-handed naming format (the letters increase going right and the numbers increase going down) while the normal format for 96-well plates is right-handed (the letters increase going right but the numbers increase going up) so the pattern needs to be reversed when inputting.

Finally, in order to improve the usability of the script and allow users not familiar with the command line to use the script effectively, a user friendly windows front end was developed in which options can be chosen using a combination of check boxes and value entry. Peaksets and calibration values can be picked from drop-down lists, and input files dragged-and-dropped to the program. If any required values are missing the program will automatically highlight the missing value rather than simply producing an error. Since its development PSAnalyse has been

used to analyse the data from a large number of experiments, including those carried out by other researchers – e.g. Dr Lea Lango-Scholey, Dr Alex Woodacre and Emma Whittle, among others. Comparison of manual analysis to the output of PSAnalyse from these experiments by myself and Dr Lea Lango-Scholey was used to verify that the output produced by PSAnalyse is accurate. PSAnalyse can be found online at http://www.jackaidley.co.uk/PSAnalyse.

4.4 Construction of a *cj1422c* knockout strain

The Δ*cj1422c*::RDH315 mutant was created in the *C. jejuni* NCTC 11168 rpsL* strain using the RDH315 insert designed to allow for a markerless mutant (section 2.4.7). However, the Δ*cj1422c*::RDH315 showed a high rate of spontaneous production of Streptomycin-resistant colonies that prevented the counter-selective step from being effectively employed, thus the markerless version was never produced and the Δ*cj1422c*::RDH315 version was used in experiments.

The mutant was created by homologous recombination using a suicide plasmid which carried the RDH315 insert together with two flanking regions homologous to regions (>450bp) of the *C. jejuni* genome adjacent to the target site. There is a large (∼1000bp) homologous region of exact sequence similarity shared between *cj1421c* and *cj1422c* which contains the poly-G tract responsible for phase variation in these two genes. Previous work has found that spontaneous recombination between these two sites can occur (McNally *et al.*, 2007) thus it was necessary to remove as much of the homologous region as possible to prevent this recombination from occurring. This homologous region, however, extends a short distance beyond the start of *cj1422c* into *cj1423c* (see figure 4.1) and so careful targeting was needed to prevent disruption to this adjacent gene.

Target	Primer	Sequence
Left	023.cj1422-FkL-FBH	ACA**GGATCC**CATTCTTTCAAATTCTTGG
	024.cj1422-FkL-RNI	GAT**GCGGCCGC**AAAAATAATAAAATATTTTATCAGC
Right	025.cj1422-FkR-FNI	TGG**GCGGCCGC**CTGAATTTGGGTTGAGC
	026.cj1422-FkR-RPI	AGC**CTGCA**GAAGGTATAAAAGAAGTAATTTTAGC

Table 4.1: Primers used in construction of *cj1422c* strain
Two primer pairs were used to amplify flanking regions either side of the target region. Left and right refer to the two flanking regions with left and right positions relative to the target in genome sequence order. Bolded part of the primer sequences are restriction enzyme cut sites for *BamHI* (GGATCC), *NotI* (GCGGCCGC), and *PstI* (CTGCA) not present in the template DNA.

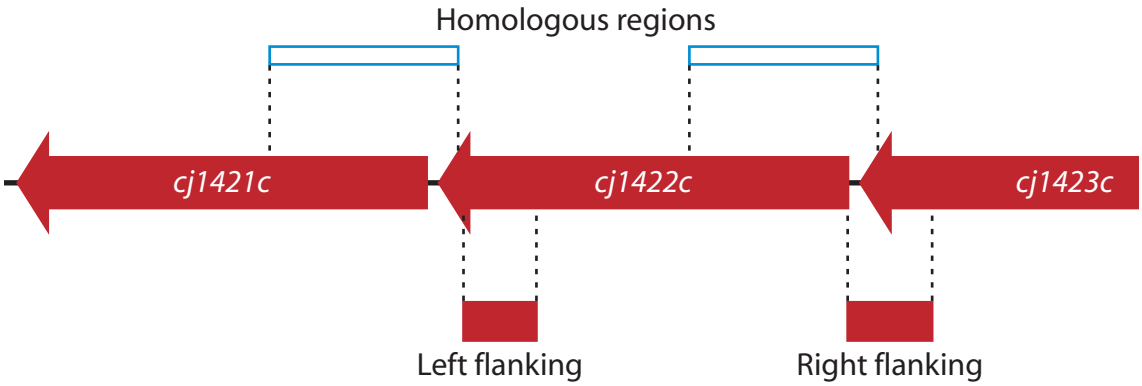

Figure 4.1: Genomic context of *cj1422c* and position of flanking regions
Figure illustrates the location of *cj1421c* and *cj1422c*, with the large homologous region shared between (top) and the position of the left (465bp) and right (602bp) flanking regions used to target the insert to *cj1422c* (bottom). Not to scale.

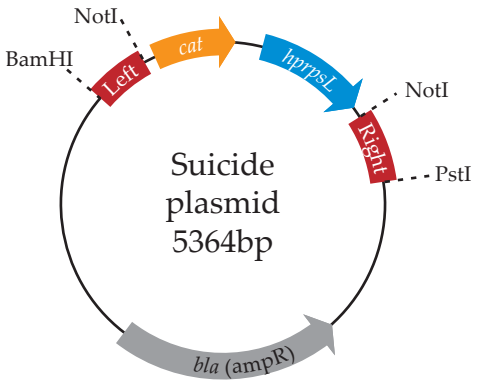

Figure 4.2: Map of suicide plasmid for the Δ*cj1422c*::RDH315 construct
Map of suicide plasmid with important restriction enzyme cut sites marked. Left and right are the flanking regions to target recombination at *cj1422c* (see figure 4.1), *cat* and *hprpsL* are parts of the RDH315 insert. The *bla* gene is the Ampicillin resistance gene present in the pUC19 vector backbone. Not to scale.

To construct the suicide plasmid (illustrated in figure 4.2), flanking sequences were amplified from genomic DNA of *C. jejuni* NCTC 11168 by Phusion PCR using the primers given in table 4.1. Restriction enzyme sites for *BamH1* and *PstI* were added to the outer ends of these flanking sequences and *NotI* sites added to the inner ends. The RDH315 cassette was cut from the pRDH315 plasmid by digestion with restriction enzyme *NotI* and then the insert separated from the remainder of the plasmid by gel purification using a commercially available kit (Zymoclean Gel DNA Purification Kit (ZYMO Research), following manufacturer's instructions). Flanking regions were attached to the RDH315 by ligation using T4 DNA ligase (section 2.4.2) before insertion in the multicloning insertion site of the pUC19 vector (Yanisch-Perron *et al.*, 1985) (New England BioLabs). The pUC19 vector and the insert were prepared for insertion by digestion with BamHI and PstI. After ligation the plasmid was transformed into chemically competent *E. coli* strain DH5α cells by heat shock (section 2.4.4) using Ampicilin (50µg/ml) and Chloramphenicol (20µg/ml) as selective antibiotics. A preparation of this plasmid was then used to transform *C. jejuni* strain NCTC 11168 rpsL* by electroporation (section 2.4.6). This insertion removes a 1340 bp section of *cj1422c*, almost completely removing the gene. *E. coli* and *C. jejuni* transformants were checked by PCR amplification and sequencing of these products. These tests demonstrated the correct insertion and deletion of *cj1422c*.

4.5 Passage with phage F336 will select for OFF expression state of *cj1421c in vitro*

Sørensen *et al.* (2011) demonstrated that *in vivo* passage of phage sensitive NCTC 11168 with phage F336 results in selection for a resistant state with *cj1421c* and *cj1422c* in the OFF state. In order to demonstrate the usefulness of phage selection for the cyclical selection assay further experiments were needed to assess the degree

and reliability of selection occurring *in vitro* and that the selection operates only on the target loci. These experiments were initially carried out in the wild type strain. The wild type strain is sensitive to phage F336 when *cj1421c* is ON and *cj1422c* is OFF and resistant to F336 in all other combinations, thus resistance can be achieved either by switching *cj1421c* to the OFF state or switching *cj1422c* to the ON state or both (Sørensen *et al.*, 2012).

Sørensen *et al.* (2012) also demonstrates that a third phase variable gene *cj1426c* produces a mild protective effect against invasion by F336, reducing the EOP by just over 50% (in contrast, changes in *cj1421c* and *cj1422c* apparently produce complete resistance). Accordingly it was necessary to determine whether this level of resistance was sufficient to produce a significant selective effect for the ON phase state of this gene.

4.5.1 Passage with phage F336 will select for phase variable expression of the phosphoramidate group

In the first experiment performed, replicates were prepared from a single inoculum grown from a frozen stock of a phage sensitive isolate with phage added to some replicates (n=5) and the rest grown without phage (n=6). 50 colonies were picked from serial dilutions of the starting inoculum (in two lots of 25 colonies from separate samples) and 20 colonies picked after a 24 hour incubation in each replicate of the experimental and control conditions. These colonies were then analysed using the GeneScan™ process and scored using PSAnalyse to determine the state of 28 phase variable tracts.

As shown in figure 4.3, there was a marked change in the phasotypes of the colonies between those isolated from the inoculum input and those from the experimental output population which was not seen in the control populations. The primary

shift observed was to the resistant state produced by switching *cj1422c* ON (to the 110 phasotype) in all five replicates. This phase change was produced by a shift in the poly-G tract of *cj1422c* from 10 to 9 Gs. There was also a slight increase in the median occurrence of the phage resistant 111 and 000 phasotypes.

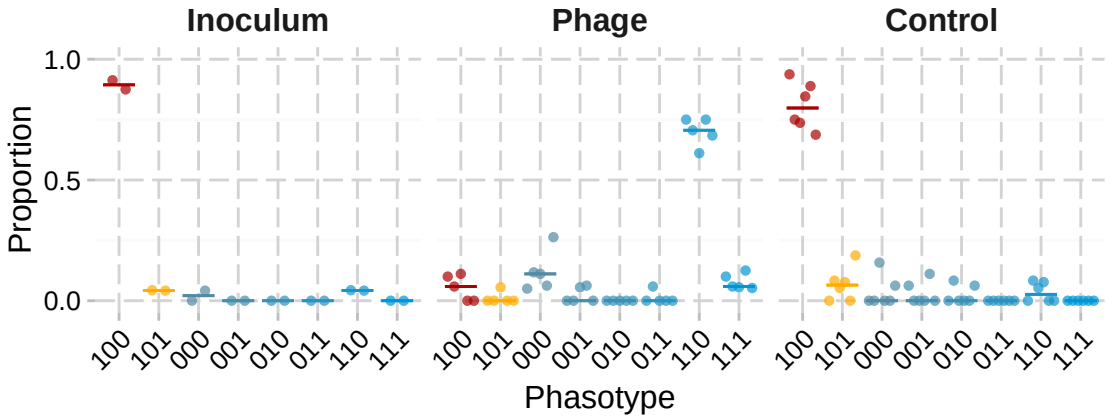

Figure 4.3: Phasotypes following F336 exposure using a common inoculum
The proportion of phasotypes (*cj1421c-cj1422c-cj1426c*) in the phage condition ($n = 5$) and control condition ($n = 6$) as well as for the inoculum ($n = 2$, sampled from the same inoculum). Points are results for each replicate whilst the horizontal bar indicates the median. Phasotypes shown in blue have shown no detectable phage susceptibility in prior experiments, whilst the two leftmost phasotypes shown in red and orange are phage susceptible.

The state of all 28 poly-G tracts amplified was also analysed (figure 4.4). Here it can be seen that there is only a marked change in the phase state of *cj1422c* and this is supported by statistical analysis. After applying a Bonferroni correction for multiple tests only the difference in observed ON expression of *cj1422c* between the experimental and inoculum conditions is statistically significant at the 5% level ($p \ll 1 \times 10^{-10}$, fitted using a generalised linear mixed model, Wald's Z).

Since this experiment showed consistent switching in *cj1422c* while the *in vivo* experiment by Sørensen *et al.* (2011) showed switching in *cj1421c* I speculated that the kind of switching observed in the output population likely depended on the relative proportions of rare variation already present in the inoculum and therefore multiple replicates drawn from the same inoculum would show similar results. Accordingly, separate inocula were seeded by picking individual colonies of the sensitive strain and then dividing each inocula to positive and negative conditions.

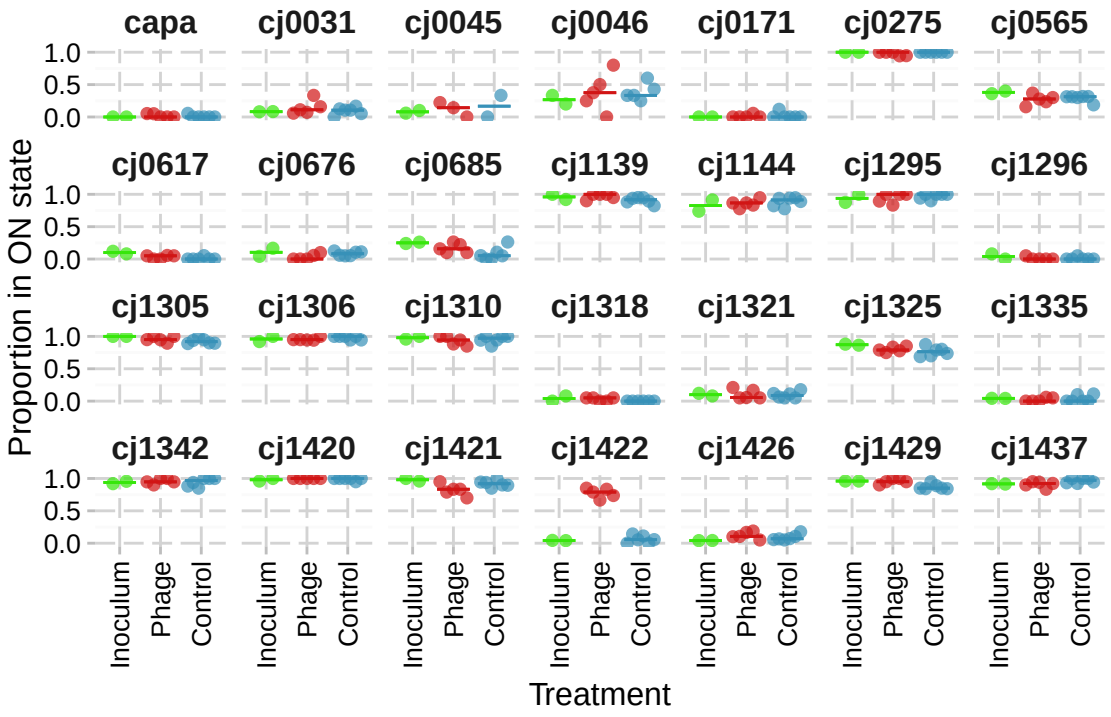

Figure 4.4: Phase state of 28 variable poly-G tracts after F336 exposure from a common inoculum

Figure shows the proportion of each poly-G tract in the ON state for each of 28 poly-G tracts. Each point indicates a replicate experiment (control $n = 6$, phage $n = 5$, inoculum shows two samples from the same inoculum), while horizontal lines show the median proportion in the ON state. Only the difference in phase state for *cj1422c* is statistically significant ($p \ll 1 \times 10^{-5}$, fitted using a generalised linear mixed model, Wald's Z).

Five separate inocula were used, with 20 colonies picked from the inoculum, the phage present and the negative control conditions.

Additionally, CFU and PFU counts were taken from each of the phage and control samples after incubation. The results show surprisingly small differences in growth between the phage present and control conditions presumably because the control condition enters stationary phase well before the incubation concludes (figure 4.5). PFU counts indicated an approximately hundred fold increase in PFU/ml demonstrating high levels of phage replication. Analysis of samples collected at the end of the control experiments confirmed that no detectable phage contamination occurred.

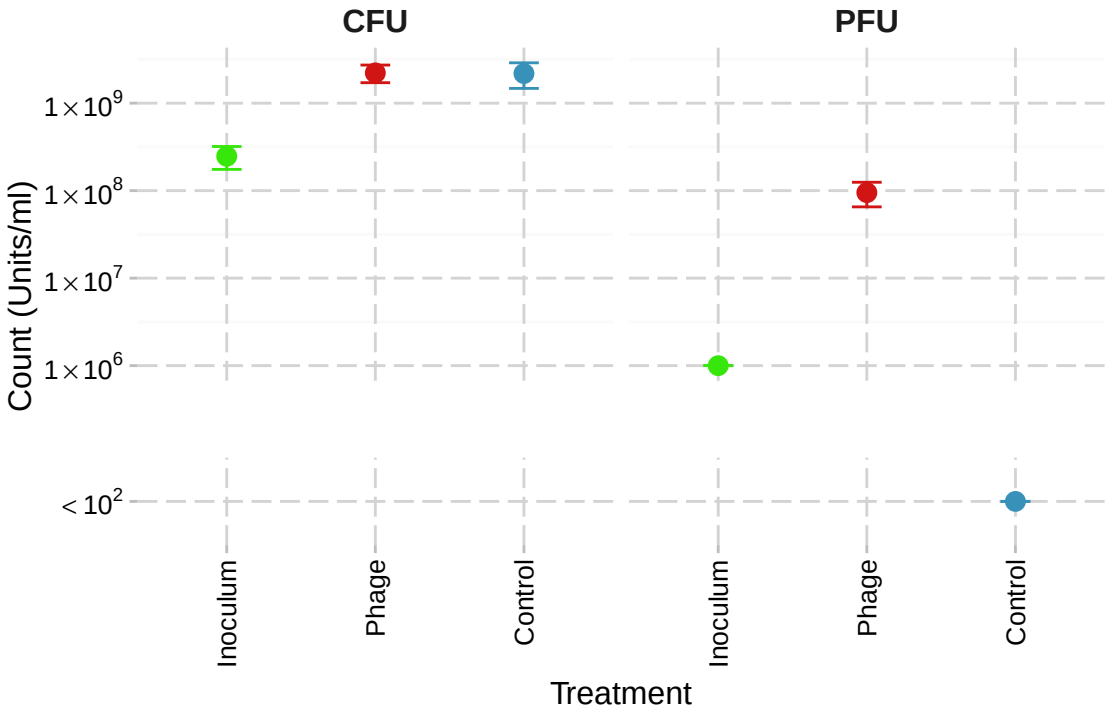

Figure 4.5: CFU and PFU counts after F336 exposure with single colony inocula
Mean CFU and PFU counts before (inoculum) and after phage-treatment and in the untreated control lines are shown ($n = 5$). Each replicate was started from a different inoculum seeded from a single colony. The minimum detectable limit was 10^2 units/ml for both CFU and PFU.

These experiments also showed a clear switching from sensitive to resistant states. The exact pattern differed between the five replicates but in each case, >90% of the output colonies were in one of the phage resistant phasotypes (figure 4.6). In 4 of

the 5 replicates, the majority of these resistant colonies had a change in the *cj1421c* tract, whilst in the final replicate there was a mixture of switching in both the *cj1421c* tract and the *cj1422c* tract resulting in a mixture of two possible resistant states. As with the previous experiment, no changes were seen in the proportions of 25 other genes, however in one replicate there was an increase in the proportion of *cj1426c* in the ON state (figure 4.7). It can also be observed that several genes showed different predominant states in the inoculum due to the bottleneck effect of starting from a single colony, and these differences are preserved in both phage and control conditions.

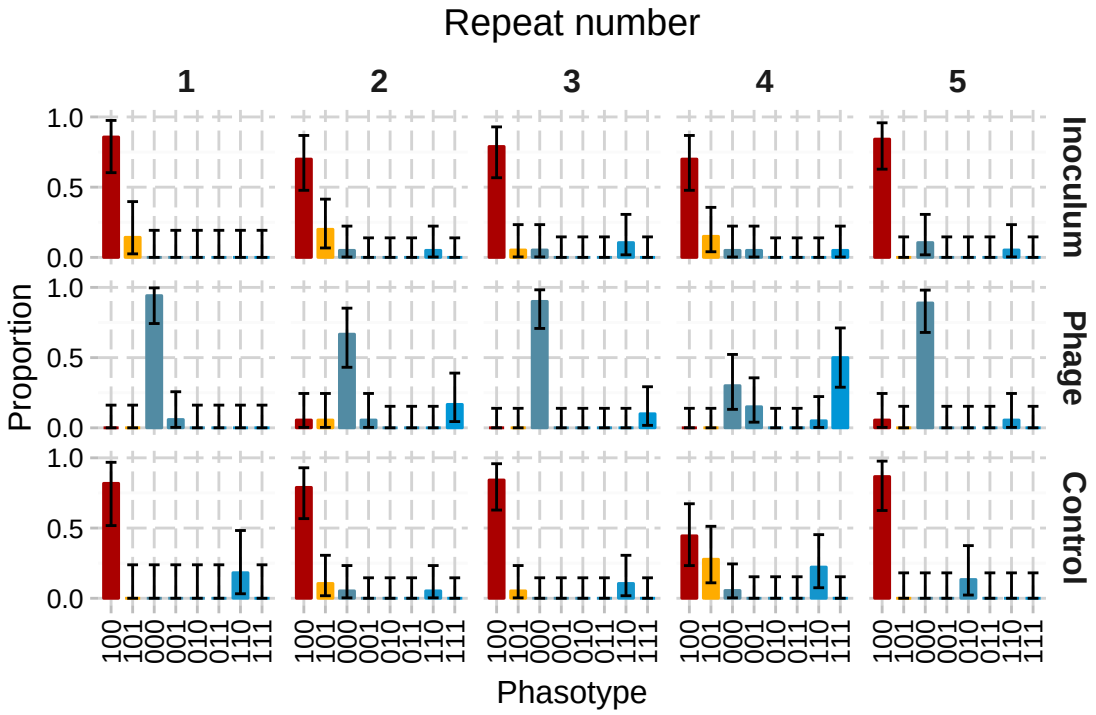

Figure 4.6: Phasotypes after F336 exposure with single colony inocula
The proportion of phasotypes (*cj1421c-cj1422c-cj1426c*) in the inoculum and phage-treated and control conditions are shown for five independent replicates starting from single colonies. Phasotypes shown in blue have shown no detectable phage susceptibility in prior experiments, whilst the two leftmost phasotypes shown in red and orange are phage susceptible. Error bars indicate a 95% confidence interval for the true proportion in each phasotype.

In the initial sensitive strain the repeat tract located in *cj1421c* has a length of 9 Gs, but this can shift to a non-functional state through either the gain or loss of a single guanine nucleotide. Changes of tract length in both directions were observed in all replicates of the phage present condition. Although the exact proportions of

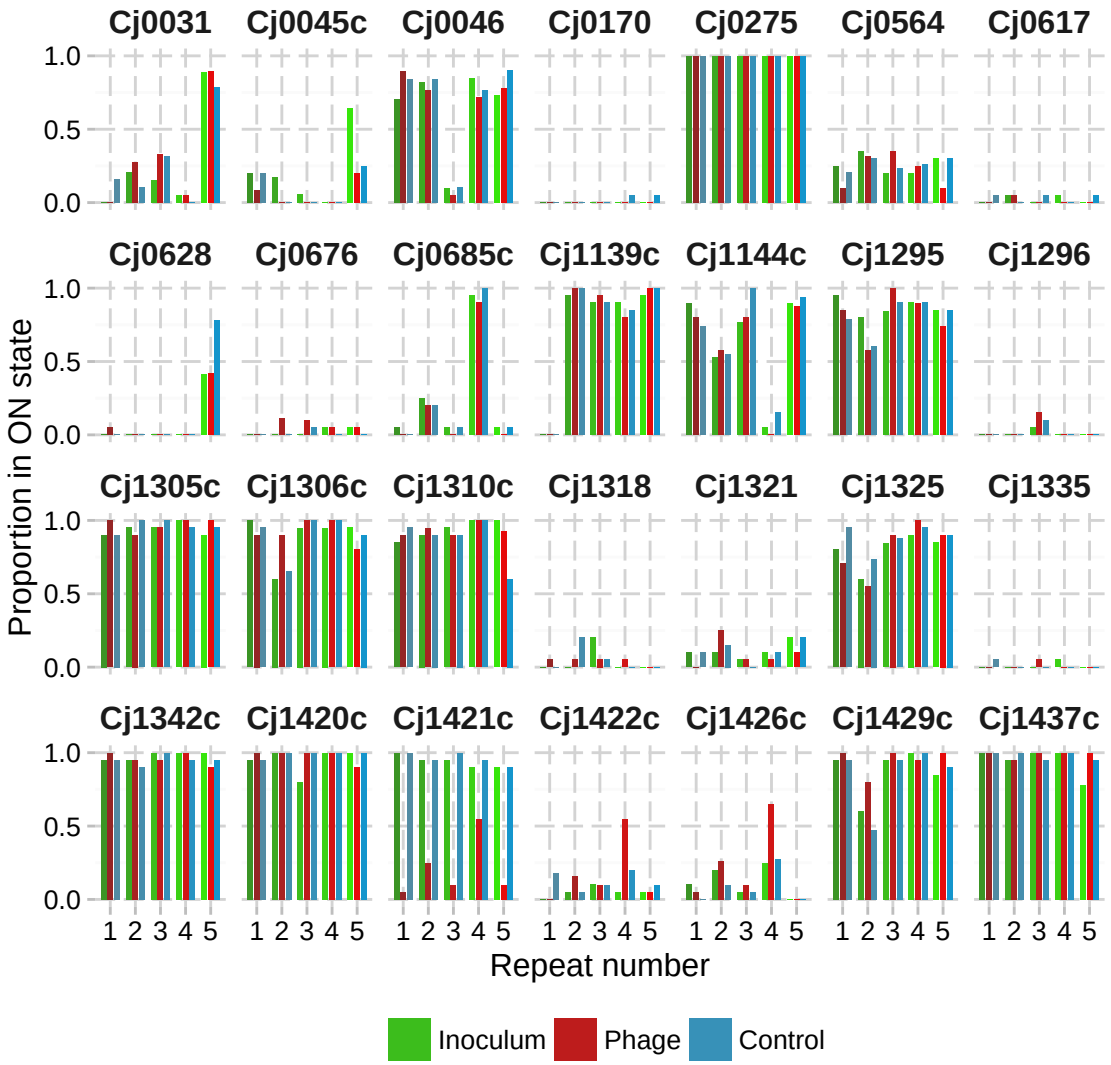

Figure 4.7: Phase state of 28 variable poly-G tracts after F336 exposure with single colony inocula

Figure shows the proportion of each poly-G tract in the ON state for 28 poly-G tracts in the inoculum, and following phage selection or control for each of 5 independent replicates, each started from a different single colony.

switching varied between replicates, the overall picture showed a slight tendency towards insertions rather than deletions (58% of observed changes were insertions), see figure 4.8. Two colonies (out of 100 in total) were recovered with the tract in the 11G state, representing two consecutive insertion events.

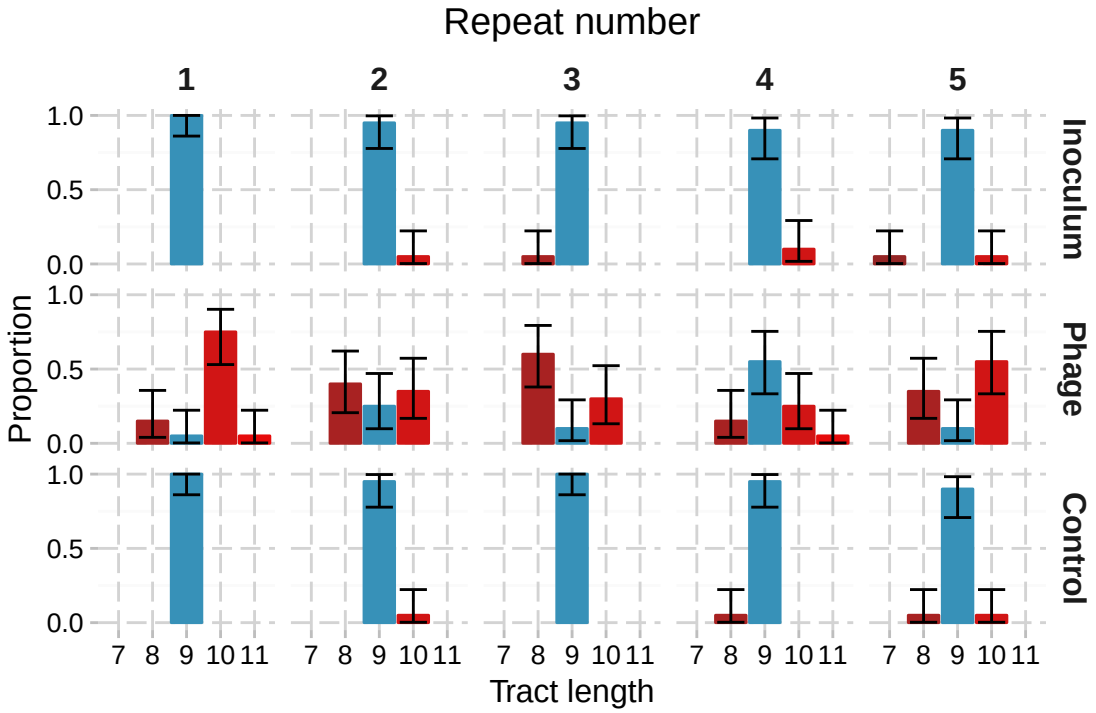

Figure 4.8: Tract lengths of *cj1421c* starting from single colony inoculum
Proportion of colonies with each tract length for gene *cj1421c* are shown for the inoculum, phage-treated, and control conditions are shown for each of 5 independent replicates starting from a single colony. The ON-state of the gene (9G) is highlighted in blue. Error bars show a 95% confidence interval for the true population proportion.

Among the inoculum and control samples, colonies with different tract lengths were far rarer (10 of 200) but the ratio of insertions to deletions was similar (60% of observed changes were insertions), no colonies showed a double insertion, one showed a double deletion (7G).

The direction of this effect is consistent with previous observations (Bayliss and Palmer, 2012) indicating that 9G tracts are more prone to insertions than deletions however the magnitude of the effect is apparently smaller than expected (that 85% of observed changes would be insertions). This may result from differences in the

experimental procedure.

Since the tract lengths for *cj1422c* began in the OFF state (10G) it required a single deletion to switch to the ON state, whilst an insertion would simply result in a different OFF state. No conversions to the longer ON state (12G) were observed and all switched colonies showed a single deletion (to 9G, data not shown).

4.5.2 Selection operates in *cj1422c* knockout strain

The results so far have all been obtained in the wild type strain, but this shows confounding selective effects on *cj1422c* whereas for the cyclical selection assay it will be necessary to concentrate the selective effect on *cj1421c*. Whilst this can be easily achieved by creating a knockout strain with *cj1422c* removed it is still necessary to demonstrate that the F336 selective effect operates as expected in this strain.

This experiment was carried out with the Δ*cj1422c*::RDH315 strain. A *cj1421c*-ON/*cj1426c*-ON isolate was obtained from this knockout strain and this isolate was then subcultured to provide the inoculum. The EOP of F336 on this strain was confirmed to be equal to the phage sensitive 101 phasotype of the wildtype strain before beginning the experiment (data not shown).

A phage selection experiment was performed using this mutant strain. The phase state of 27 tracts (since *cj1422c* has been removed) are shown in figure 4.9 for the inoculum and three replicates of the control and experimental conditions. After passage with phage, 100% of the recovered colonies had *cj1421c* in an OFF state, with the majority (50 of 59 colonies) at 10G, 2 at 11G and 7 at 8G.

As before there were no noticeable differences between the inoculum and the two experimental conditions in any other SSR tract except for the target gene, *cj1421c*.

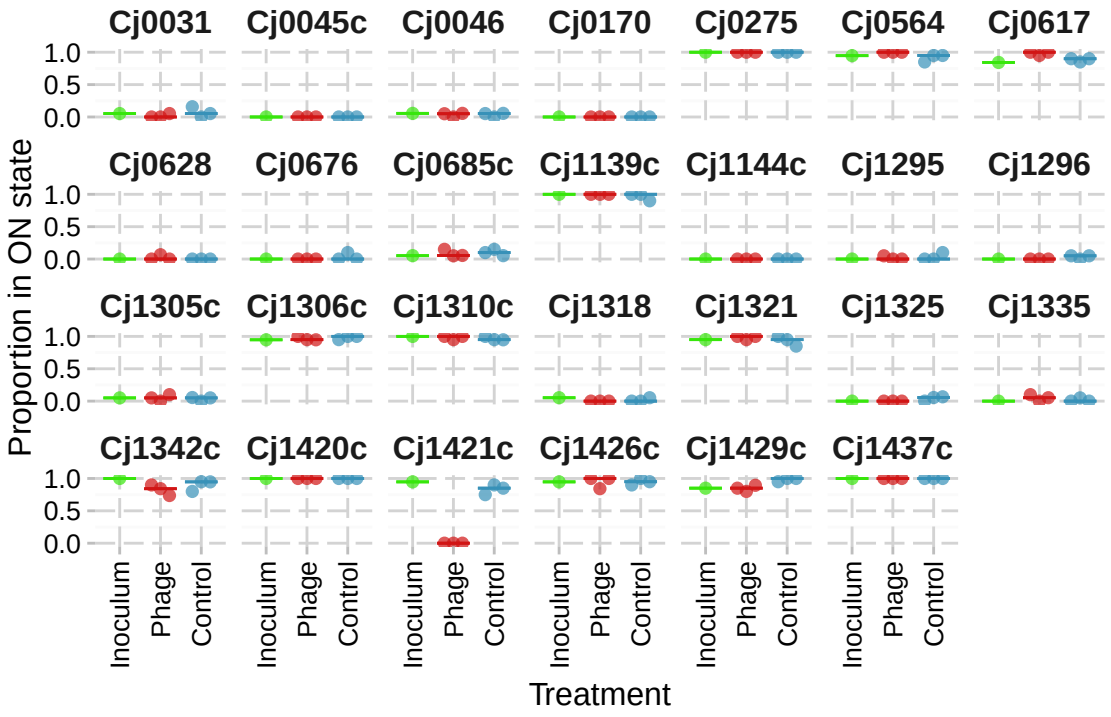

Figure 4.9: Phase state of 28 variable poly-G tracts after F336 exposure in *cj1422c* knockout

Figure shows the proportion of each poly-G tract in the ON state for each of the 27 remaining poly-G tracts in the Δ*cj1422c*::RDH315 knockout strain. Each point indicates a replicate experiment (control $n = 3$, phage $n = 3$), while horizontal lines show the median proportion in the ON state. Only the difference in phase state for *cj1421c* is statistically significant ($p \ll 1 \times 10^{-10}$, fitted using a generalised linear mixed model, Wald's Z).

The near total switching to the OFF state of *cj1421c* in the experimental condition is highly significant ($p \ll 1 \times 10^{-10}$, fitted using a generalised linear mixed model, Wald's Z) after applying a Bonferroni correction for multiple testing while no other change is statistically significant at the 5% level.

4.6 Passage with human sera will select for the ON state of *cj1421c*

The results from van Alphen *et al*. (2014) indicate that the MeO*P*N group is protective against human sera, however this result was obtained in the 81-176 strain and no distinction was made between the MeO*P*N groups added by Cj1421c and Cj1422c. Thus it was necessary to carry out experiments to show that (1) the MeO*P*N group is protective against killing by serum in the NCTC 11168 strain; (2) that the group added by Cj1421c produces the effect; (3) that the effect is strong enough to select for a population with *cj1421c* almost entirely in the ON state when starting from a population mostly in the OFF state; and (4) that no other phase variable tracts are selected for by the sera.

4.6.1 Incubation with human sera selects for expression of one of the MeO*P*N groups

In order to initially demonstrate the viability of the technique an assay was carried out with pooled human sera collected from healthy human subjects (Bayliss, personal communication). Initially, 10% sera was used and samples were collected after 30 and 60 minutes and compared to samples taken from the inoculum which was prepared from a isolate previously determined to have both *cj1421c* and *cj1422c*

in the OFF state with the tract in *cj1421c* being 8Gs long and the tract in *cj1422c* being 10Gs long. Since the stability of the tracts is dependent on the tract length, it is expected that *cj1422c* will show higher rates of phase variability than *cj1421c* in this isolate. As shown in figure 4.10 a strong selective effect for the ON state of *cj1422c* was seen after 60 minutes but not after 30 minutes; a slight rise in the proportion of *cj1421c* in the ON state was also observed but this was not statistically significant. After 60 minutes the selection had produced a change in *cj1422c* from 7% ON to 76% ON and an overall change in expression of one or more MeO*P*N transferase genes from 10% in the inoculum to 90% after 60 minutes. After 60 minutes no change was observed in eight other phase variable tracts analysed confirming that the selective effective was localised to the two MeO*P*N transferase genes.

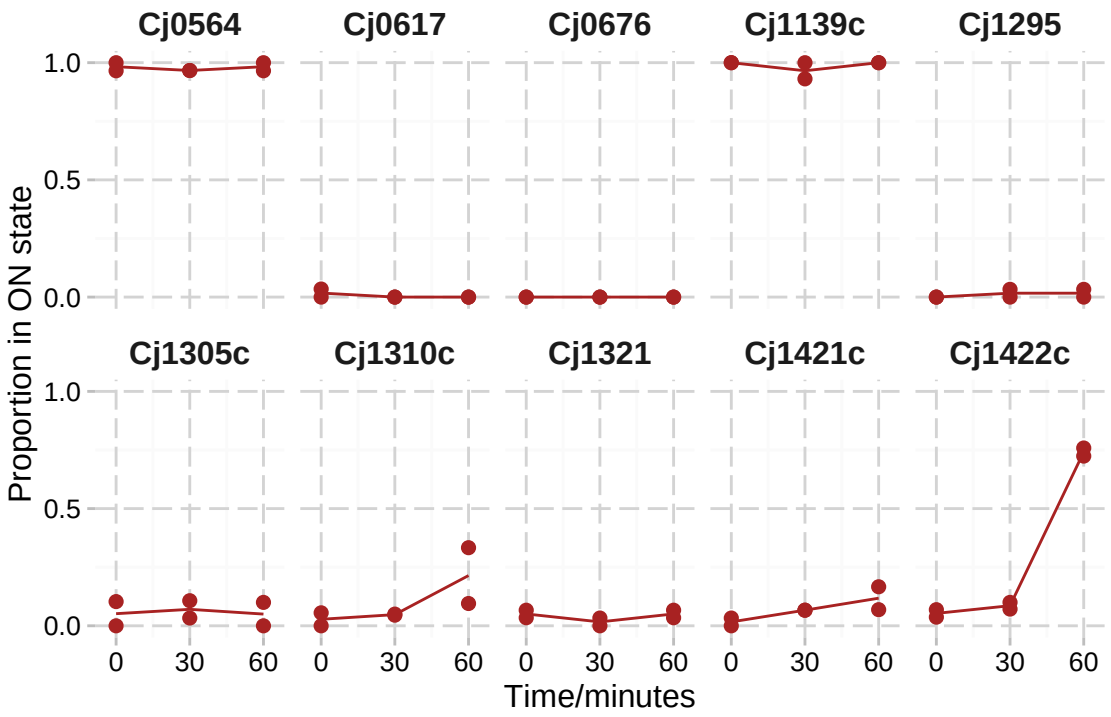

Figure 4.10: Phase state of 10 variable poly-G tracts during incubation with pooled human sera
Figure shows the proportion of each poly-G tract in the ON state for 10 of the poly-G tracts at 0, 30 and 60 minute timepoints during incubation with 10% pooled human sera. The experiment was carried out in duplicate and values for both runs are shown. 20 colonies were picked at each point to determine proportion ON.

4.6.2 Populations enriched for either *cj1421c*-ON or *cj1422c*-ON show higher survival under serum selection

The previous experiment showed a strong bias towards switching to *cj1422c*-ON rather than *cj1421c*-ON. The likely explanation for this is the difference in switching rates. However, there was also a potential difference in protective effects of the two MeO*P*N modifications against sera. To test this isolates with either *cj1421c*-ON or *cj1422c*-ON and the other gene in the OFF state were subcultured and then incubated with 10% sera in comparison to the isolate with both genes in the OFF state. CFU counts taken after 60 minutes show that populations enriched for either *cj1421c*-ON or *cj1422c*-ON show over an order of magnitude higher survival rates than the unenriched population and that survival rates are similar between the *cj1421c*-ON and *cj1422c*-ON enriched populations (figure 4.11). These differences were not apparent in the heat inactivated sera control.

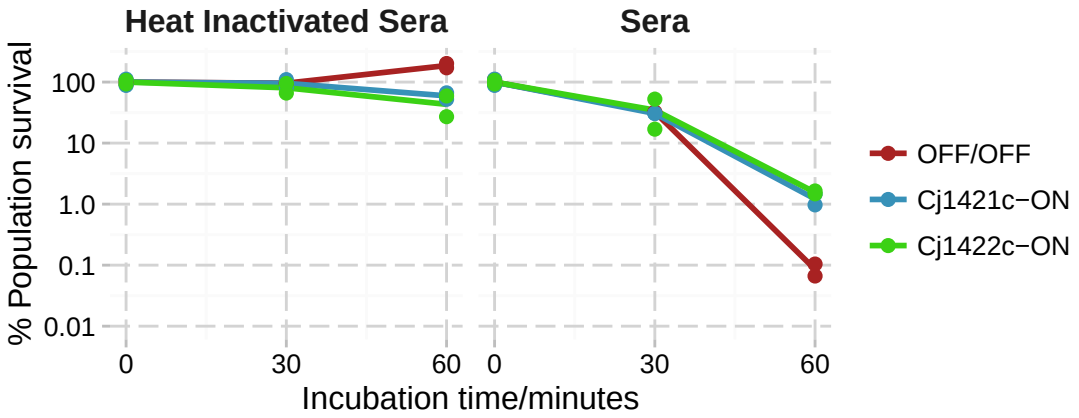

Figure 4.11: Survival of populations of different initial phasotype in pooled human sera
CFU/ml measurements were taken from the inoculum (0 minutes), after 30 minutes and after an hour of incubation with either 10% pooled human sera or heat inactivated sera for populations starting with both MeO*P*N transferases in the OFF state and with either *cj1421c* or *cj1422c* in the ON state and the other in the OFF state. In each case, the CFU counts have been normalised to the initial measurement to give the %age survival data shown on the graph.

4.6.3 The strength of serum selection effect varies between individual sera samples and pooled serum samples

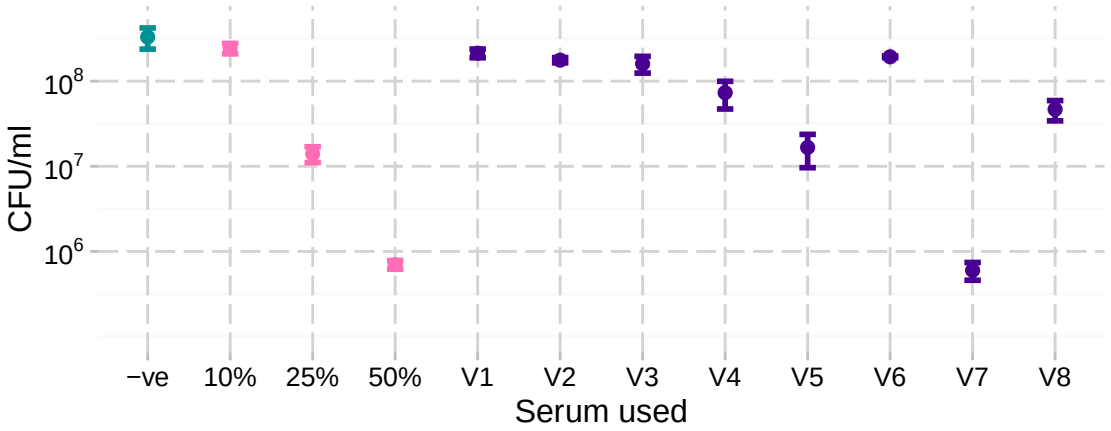

Figure 4.12: Bacterial survival after incubation in pooled serum II or serum samples from individual volunteers
Figure shows recovered colony counts after 60 minutes incubation with pooled serum or serum from individual volunteers. Negative control used 50% pooled serum inactivated by 45 minutes heating to 56°C. 10%, 25%, and 50% refer to concentrations of pooled serum II used, whilst V1-V8 refer to the eight individual serum samples pooled into pooled serum II, all used at 50% concentration. Error bars show standard deviation of three measurements of the colony count from the same experiment.

Because more of the pooled human serum used in these experiments (serum I) was unavailable for use, new pooled serum was collected from eight healthy human volunteers (serum II). In contrast to serum I experiments carried out at 10% serum with serum II for 60 minutes failed to show a selective effect (data not shown). To investigate the basis for this difference in efficacy the selective effect produced by serum samples from each of the eight volunteers included in the serum II pool was investigated, alongside increased concentrations of pooled serum II. These experiments were carried out as before, but with increased concentration of serum. 50% serum was used with each individual serum sample (V1-V8). As shown in figure 4.12, increasing the concentration of serum recovered the killing effect, whilst marked differences in killing effect were observed between individuals. Individual colonies were picked from output population after passage with serum from volunteers V5, V7 and V8 (as these showed a killing effect) as well as from the

25% and 50% pooled serum samples and analysed to determine the state of genes *cj1421c* and *cj1422c*, as shown in figure 4.13. Only 50% pooled serum II or serum from volunteer V7 showed a selective sweep for expression of one or more of these genes, that is that they show the almost total loss of the 00 (*cj1421c*-OFF/*cj1422c*-OFF) phasotype by phasotypes with either *cj1421c*-ON (10) or *cj1422c*-ON (01) or both (11).

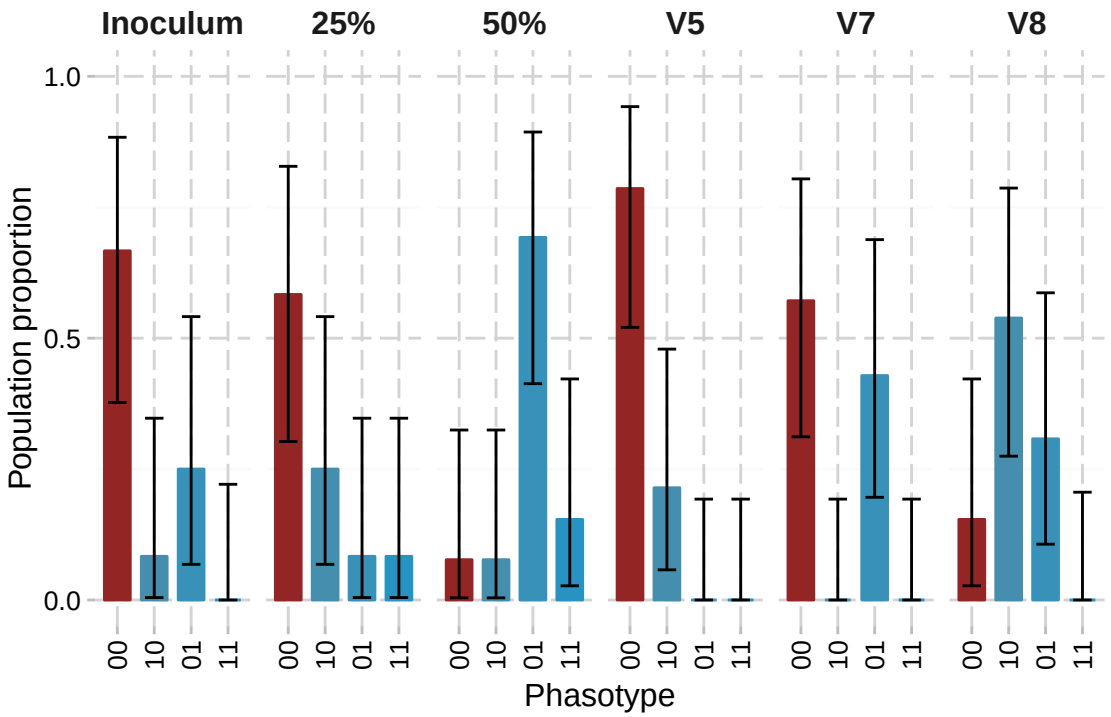

Figure 4.13: Selective effect of pooled serum II concentrations and selected individual volunteers
Figures shows proportion in each of four possible phasotypes of *cj1421c* and *cj1422c* with either 25% or 50% pooled serum II or 50% of each of three individual serum samples (V5, V7, and V8). Proportions are based on sample of 12-14 colonies. Error bars show 95% confidence interval for proportions calculated using the midP adaptation of the Clopper-Pearson interval.

4.7 Both the ON and OFF phase states of *cj1421c* can be selected for independently of other genes

The results presented in this chapter indicate that it is possible to employ phage F336 as a selective agent to take a population of *C. jejuni* with *cj1421c* primarily in

the ON state and recover a population primarily in the OFF state after selection using a Δ*cj1422c* strain. No other PV loci was affected by this selective assay. Conversely, incubation of a population primarily with the two MeO*P*N transferases in the OFF state with pooled human serum allows recovery of a population primarily composed of cells with one or other of the MeO*P*N transferases in the ON state. Moreover, there is no evidence of selection for changes in any other phase variable loci. The selective effect of pooled serum is variable between samples, with the selective effect of pooled human serum II only apparent at concentrations of 50% and there is some evidence of variation between serum samples collected from individual volunteers. Thus, taken together, these data indicate that these two forms of selection should permit the performance of a cyclical selection assay operating on *cj1421c* in a Δ*cj1422c* strain.

Chapter 5

A cyclical selection assay

5.1 Summary

In this chapter, I combine the phage and serum selection methods described in chapter 4 to produce a complete ON→OFF→ON cycle of *cj1421c* in a Δ*cj1422c* strain of NCTC 11168 thus demonstrating that a full cycle is possible. This cycle was carried out with pooled serum I and shows no impact on genes other than *cj1421c*. However, unlike pooled serum I, pooled serum II does not remove phage during the serum selection step and thus the cycle cannot proceed. I further demonstrate that phage cannot be effectively removed by repeated washing or by growth in media which is not cation enriched (i.e. substituting BHI for CBHI).

5.2 Introduction

In section 4.5 of the last chapter it was demonstrated that phage F336 will select for the OFF state of *cj1421c*, and in section section 4.6 that pooled human serum will

effectively select for the ON state of the two MeOPN transferases. The next step is to combine these selective assays together to create a complete cycle starting from a population of NCTC 11168 Δ*cj1422c*::RDH315 with *cj1421c* mostly in the ON state, selecting with phage F336 to recover a population mostly in the OFF state, and then applying selection with pooled human serum to return to a population with *cj1421c* mostly in the ON state (see figure 5.1). This chapter aims to demonstrate that these two selective agents can be combined in a viable cyclical selection assay and that no PV loci other than *cj1421c* are impacted by this cycle.

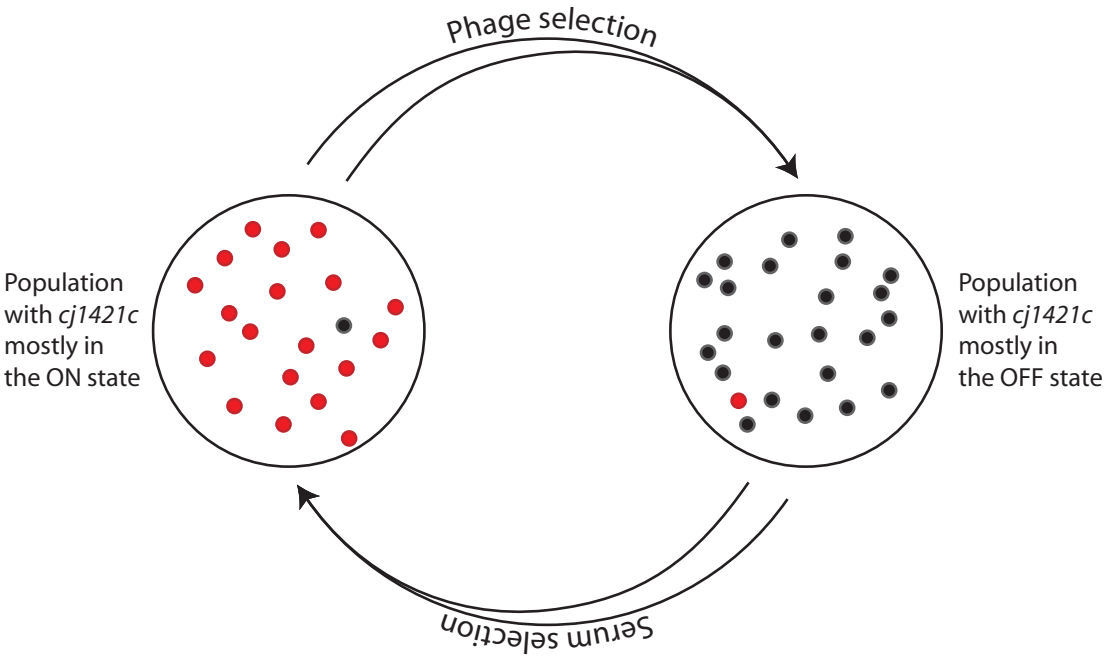

Figure 5.1: Design of the cyclical assay
Diagram illustrating the design of the cyclical assay. From an initial population with *cj1421c* almost entirely in the ON state, selection with phage F336 is applied to produce a population almost entirely in the OFF state. Then the cycle is completed by taking this population and incubating it with pooled human serum to select for a population almost entirely in the ON state again. This cycle can then be repeated.

5.3 A complete cycle is possible

For a trial cycle of the cyclical assay three experimental replicates, and three controls, were seeded from a common inoculum grown overnight in CBHI from a

re-suspended sweep of NCTC 11168 Δ*cj1422c*::RDH315/*cj1421c*-ON grown on an MHA plate from a frozen -80°C stock. This inoculum was adjusted to an OD_{600} of 0.01 (approximately 10^7 CFU/ml) before use. The inoculum was serially diluted to obtain accurate CFU/ml and colonies to analyse with the 28-locus-CJ11168 PV-analysis assay (see section 2.3). Each replicate was passaged in 10ml volumes in the presence of F336 phage with a starting MOI of approximately 0.01 for 22 hours at 42°C under microaerobic conditions. SM buffer was added in place of the phage suspension in negative controls. After completion of the phage selection cycle, a 900µl sample was taken from this culture and directly mixed with 100µl of pooled serum I (for a final 10% serum concentration). This mixture was incubated at 37°C for 60 minutes in a sealed 1.5ml microcentrifuge tube. The negative control was performed by substituting active serum for serum inactivated by heating to 56°C for 45 minutes (note that this is not a true negative control because of the continued presence of phage in the solution). Samples were collected after the phage selection step from all three replicates and three negative controls, and after the serum selection step from all three replicates and three negative controls. These samples were serially diluted to determine CFU/ml and to allow colonies to be picked for the 28-locus-CJ11168 PV-analysis assay. Further samples were centrifuged to pellet out the bacteria and the supernatant collected for phage titration in order to determine the phage concentration (PFU/ml) at each stage and ensure no contamination of the phage-negative controls.

Starting CFU/ml was measured at 4.7×10^7, and starting PFU/ml was 2×10^5. Figure 5.2 shows the state of the 27 remaining PV loci at each stage of the experiment, demonstrating the successful completion of one complete cycle with no impact on any PV loci other than the target gene *cj1421c* ($p \ll 1 \times 10^{-5}$, fitted using a generalised linear mixed model, Wald's Z). PFU/ml reached 5×10^7 after the phage selection step, but was reduced to undetectable levels ($< 10^2$ PFU/ml) after the serum selection step. This reduction was not seen in the heat inactivated serum control. The bacterial population reached 3×10^8 CFU/ml after phage selection

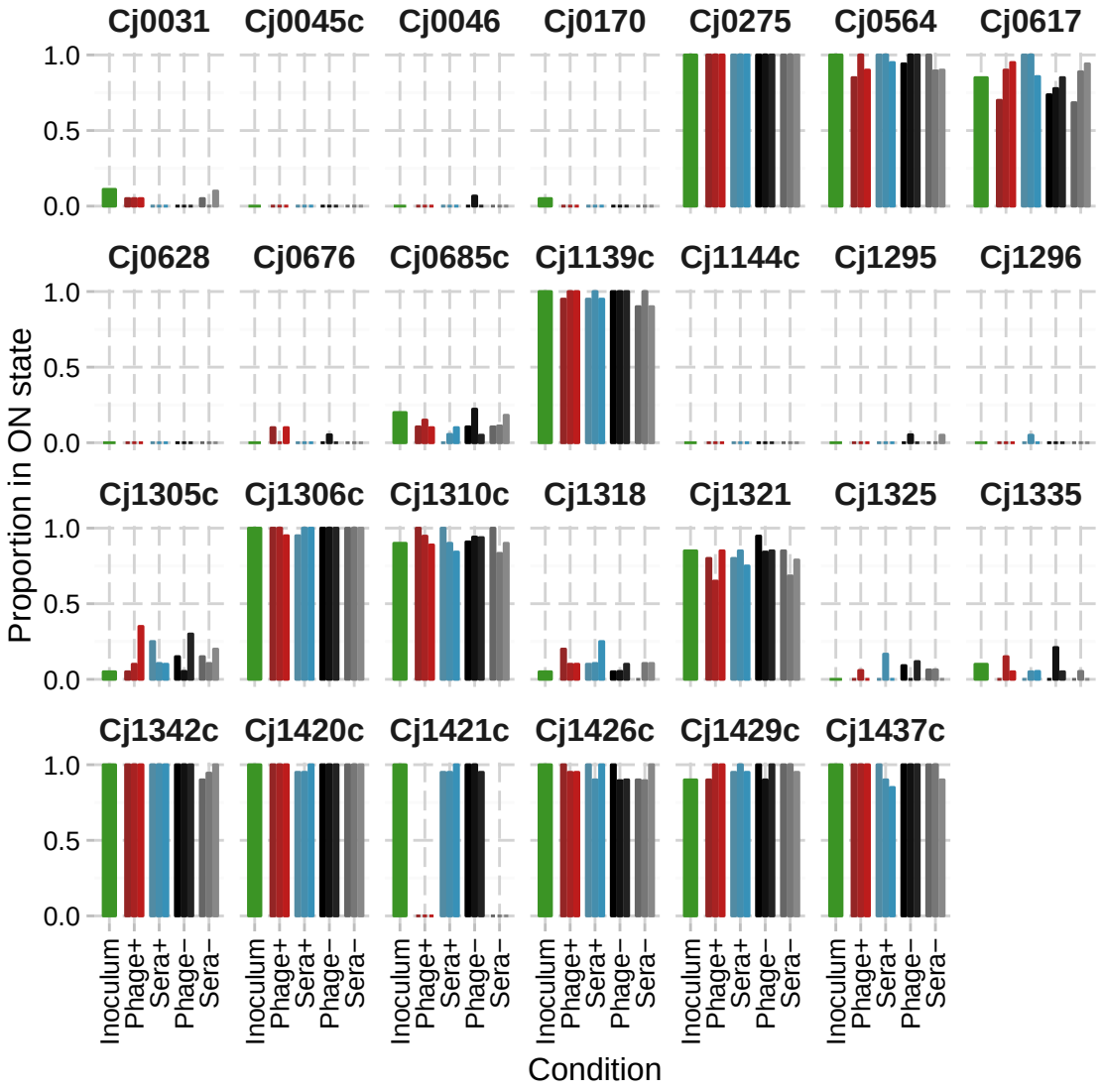

Figure 5.2: State of 27 phase variable genes through one cycle of the cyclical assay
Figure shows proportion of each PV gene in the ON state at each stage of a complete cycle of the cyclical assay. Bars are shown separately for each replicate. Each proportion was determined from 20 colonies, with valid points for ≥ 16 samples for every condition and replicate. Only the changes in *cj1421c* between inoculum and post-phage selection, and between post-phage and post-serum selection are statistically significant ($p \ll 1 \times 10^{-5}$ for both, fitted using a generalised linear mixed model with logit link function, Wald's Z).

but was reduced below 10^5 CFU/ml after serum selection. This indicates the likely need for a recovery period between cycles of the assay in order for the population to reach sufficient size for the phage selection step of the next cycle to proceed.

5.4 Not all pooled human serum samples remove phage

Unlike pooled serum I, pooled serum II does not effectively reduce PFU/ml from the levels present after the phage selective step to undetectable levels even after incubation for 60 minutes with 50% serum at 37°C. This prevents the completion of the selective cycle when using this serum as phage is constantly present through the cycle.

5.4.1 Repeated washing does not effectively remove phage

Phage F336 was added to a liquid culture of the phage sensitive NCTC 11168 adjusted to an OD_{600} of 0.01 at an approximate MOI of 0.01. After 24 hours of growth in CBHI with shaking, PFU/ml reached approximately 10^8. Washing in CBHI by removing the supernatant and replacing with fresh CBHI reduced the PFU/ml so that it was detectable at dilutions of up to 10^{-5} but four further washes failed to significantly further reduce the PFU/ml. Since even with a low MOI phage will rapidly proliferate this is insufficient to effectively remove the phage from the population and allow phage-free selection for *cj1421c*-ON.

5.4.2 Growth in media without cation enrichment does not significantly inhibit phage proliferation

The phage growth protocol from Sørensen *et al*. (2011) calls for enrichment of BHI with additional Mg^{2+} and Ca^{2+} ions to produce CBHI. It was therefore suspected that using media without enrichment would inhibit the proliferation of phage F336 and thus phage could be limited by re-suspending in un-enriched media after the serum selection step. However, replacing CBHI with either BHI or MHB and growing as above did not significantly alter the PFU/ml after 24 hours and the recovered PFU/ml was 10^7-10^8 in all replicates.

5.5 Creating a cyclical selection assay is possible

The results presented in this chapter demonstrate that phage and human serum selection can be combined to produce a complete cycle of selection without altering the proportion of ON variants in other PV loci. It is possible to start with a population of *C. jejuni* strain NCTC 11168 Δ*cj1422c*::RDH315 with *cj1421c* mostly in the ON state, grow it in the presence of phage F336 to recover a population with *cj1421c* mostly in the OFF state, and then by incubating with pooled human serum recover a population mostly in the ON state similar to the starting population. However, this cycle relies on the phage removal property of the pooled serum and this property is not reliably present in pooled human serum samples. Thus for the cycle to work either a pooled human serum that removes phage must be used, or an alternative phage removal method found.

Chapter 6

In silico modelling of selective and non-selective bottlenecks

6.1 Summary

In this chapter I move onto investigating the behaviour of the cyclical selection in an *in silico* simulation. Before investigating the behaviour of phase variable populations under selection, this chapter first includes a simulation of non-selective bottlenecks. Both simulations are based on stochastic modelling of phase variation in growing populations. The non-selective bottleneck data show that bottleneck size has a qualitative effect on the outcome, with small bottlenecks producing stochastic changes in major phasotypes while large bottlenecks preserve diversity. The selective bottleneck model indicates that phase variation is favoured across a broad range of conditions while the favoured tract length (and, therefore, PV rate) is typically determined by the number of generations occurring in a stable

environment independently of the strength of selection. However, the exact response depends on the selection of ON length and the way that mutational rate parameters are derived from experimental evidence. Simulations based on the empirical data from chapter 4 indicate that experimentally plausible conditions will favour phase variable strains over non-phase variable constructs but that all available conditions are likely to favour the same phase variable tract length, G9 in the wild type *cj1421c* gene. It may be possible to create constructs with a different ON-length, i.e G10 or G11, with alternate dynamics.

6.2 Introduction

In this chapter I present two *in silico* models of phase variation. The first looks at the behaviour of phase variable populations under non-selective bottlenecks; the second looks at the effect of alternating selective bottlenecks. This second model looks at two aspects of the behaviour of a single phase variable tract under alternating conditions that exert selection for either the ON or OFF state of the tract, firstly in general and secondly parametrised by the results described in chapter 4. The model used to predict the outcome of potential experiments using the cyclical assay described in chapter 5. These models share some of the same implementation and the source code for both can be found online at `http://www.jackaidley.co.uk/pv-models`.

Pathogenic bacteria, in general, are subject to acute population reduction during their lifecycle. These may result from transmission from host to host, or within hosts from compartment to compartment. Experimental evidence for this comes from the use of isogenic, tagged mutants in experimentally induced bacterial pathogenesis and suggests that these bottlenecks can be as acute as a single cell (Moxon and Murphy, 1978, Gerlini *et al.*, 2014). These bottlenecks can severely

reduce population diversity (Wahl *et al.*, 2002), or result in the preservation of phenotypes adapted to higher rates of transmission despite lower general fitness (Handel and Bennett, 2008). The potential of phase variation to rapidly generate large degrees of phenotypic variation may act counter to the reduction imposed by these bottlenecks. Additionally, bottlenecks imposing selective effects will act as non-selective bottlenecks on loci not involved in the selection. This is particularly relevant to PV loci because of the high mutation rate of these genes.

Experiments carried out using reporter constructs, or antibody colony blotting, indicate that the phase variation rate of PV loci in *C. jejuni* is dependent upon the length of the poly-G tract (Bayliss *et al.*, 2012), and one of the primary aims behind the development of the cyclical assay described in chapter 5 is to study whether these differences in phase variation rate have biological and evolutionary significance. The second model seeks to address this question and simulate the conditions produced by the selective forces described in chapter 4, as well as produce data on expected tract length distributions that can be compared to the results in chapter 3. The behaviour of phase variable tract length under alternating selection has previously been addressed by Palmer *et al.* (2013) based on mutation rates in *H. influenzae*. The model described here differs from this earlier work in two aspects: (1) it uses a variable population size, and (2) more importantly, the patterns of *C. jejuni* mutation show significant differences from the *H. influenzae* model. In *H. influenzae* the mutation rate increases with tract length in a linear fashion whereas the relationship is more complex in *C. jejuni*. In *H. influenzae* deletions are always more likely than insertions over the studied range whereas in *C. jejuni* the ratio depends on the tract length with shorter tracts favouring insertions and longer tracts favouring deletions. These differences may produce qualitative differences in the behaviour of the model.

6.3 The impact of non-selective bottlenecks on phase variable populations

6.3.1 Defining the problem

Where input and output populations differ there are three possible explanations of how this could have occurred: selection, genetic drift, and the imposition of a non-selective bottleneck or bottlenecks (Aidley and Bayliss, 2014). In selection, these changes result from differential fitness between forms favouring one over another and thus changing population ratios over time; in genetic drift, the changes simply result from stochastic effects; and in a non-selective bottleneck the changes result from sampling effects where the population structure after the bottleneck does not match that prior to the bottleneck in a manner akin to the "founder effect" (Barton and Charlesworth, 1984). In this section, I address the third of these possibilities, which could potentially produce significant changes in population structure of PV expression states. Pre-existing experimental data was available and provided an opportunity to compare the results of a new *in silico* model to the results of *in vitro* experiments. These experiments (performed by Shweta Rajopadhye and Nwanekka M. Akinyemi; Bayliss, personal communication) involved an *in vitro* investigation of non-selective bottlenecks in which populations were grown on plates and serially transferred with bottlenecks of various sizes. Colonies from the input and output populations were collected and analysed to ascertain the state of the phase variable genes in the these populations. The *in silico* model, therefore, was designed so that it could model these experiments.

To model this situation, the simulation needs to model the behaviour of phase variable genes in a growing population with serial bottlenecks.

6.3.2 Simulating a growing population

The poly-G tracts in *C. jejuni* exhibit a range of lengths and the rates of PV of each locus depend on the numbers of repeats in the tract and the direction of switching (Bayliss *et al.*, 2012). In order to simplify the *in silico* simulation of multiple PV loci, the switching rates for each locus were reduced to a binary ON/OFF model with symmetrical switching rates for ON-to-OFF and OFF-to-ON directional switches. To capture the dynamics of growth and bottlenecks, the model has a discrete growing population which doubles in size at each generation and has a probability of switching applied at each generation (figure 6.1(a)). This switching is applied to the new growth so one half of the new generation is identical to the previous generation while the other half is subject to a chance of changing phasotype. This mimics the expectation from slipped-strand mispairing of an SSR where the template strand is unaltered while the newly-synthesized strand contains an indel (c.f. figure 1.1). In order to study the behaviour of phasotypes (i.e. combinations of expression states for multiple phase-variable genes), the model incorporated multiple PV genes, each of which switches independently. In most simulations, six independently varying genes were simulated. This number was chosen to allow simulations to be performed in a reasonable period of time since simulation times increases approximately four-fold with each additional gene. Additionally, the number of colonies picked in the *in vitro* experiments (30) limits the number of phasotypes that can reasonably be quantified, and so limiting the simulation to six genes ensures that the results from *in silico* and *in vitro* experiments can be compared.

To increase performance the population was modelled as counts of each phasotype rather than individual cells in a manner mathematically equivalent to modelling each cell (figure 6.1(b)). To do this, the change of phasotype was generated by picking random numbers from the multinomial distribution. The binary representation of this number represents the digits of the phasotype which will change –

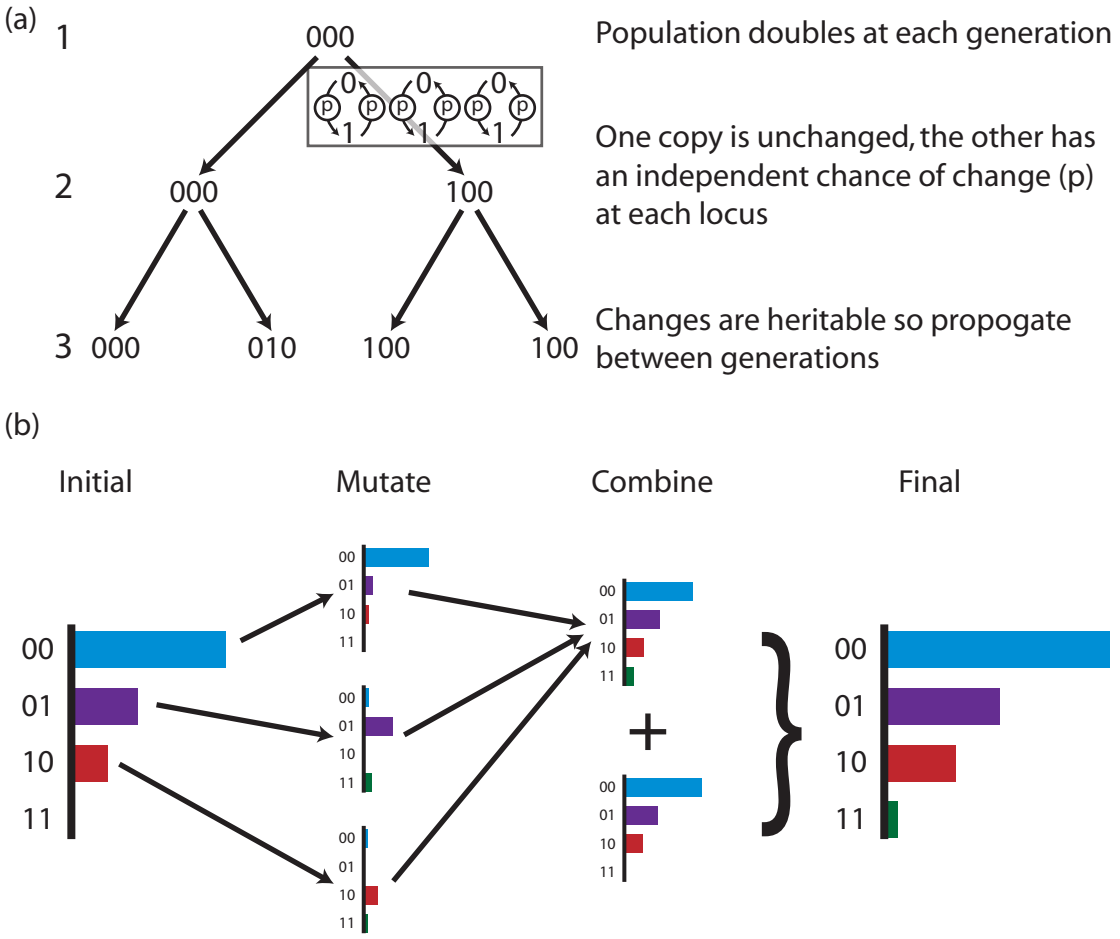

Figure 6.1: Simplified diagram of simulation method
(a) Simulation models a growing population of phase variable cells. At each generation the population doubles, with half of the cells retaining the phasotype of the parent cell and the other half subject to a fixed probability of mutation at each phase variable loci (3 shown, 6 in most simulations). These mutations are propagated through subsequent generations so early mutations have a larger impact than later mutations. (b) In order to increase the speed of the simulation, cells are modelled as populations of each phasotype (for the purposes of the diagram just 2 phase variable genes and thus 4 phasotypes are shown). Each generation, the population doubles with half of this population carried forward unchanged, and half subject to mutation. Mutations are determined randomly for each phasotype by sampling the multinomial distribution. Note that although changes at multiple loci can occur they are extremely rare. These subpopulations are combined and added to the initial population to produce the final population.

i.e. `010000` (2 in decimal) would represent a change of state in the second gene, and `001010` (20 in decimal) would represent change of state in the third and fifth genes. This is trivially and efficiently implemented by using the exclusive OR operator (see section 3.3 for a description of this operator) with the change number and the existing phasotype as the inputs. In order to exactly simulate independent switching at each of the six genes, a number between 0 and 63 (with 0 representing no change, and 63 change at all loci) was derived from the multinomial distribution with the probability of each number being based on its binary representation. Summing the digits in the binary representation gives the number of loci which will change from the original phasotype, and thus the probability for each number can be given as $P(n) = p^b$ for $n \neq 0$ where b is sum of the digits in the binary representation of n. For example, $P(1) = p$ and $P(4) = p$ (as 1 represents phasotype `100000`, and 4 represents `001000`) but $P(5) = p^2$ and $P(63) = p^6$ (as 5 represents `101000`, and 63 represents `111111`). p^6 is such a vanishingly small number that these results effectively do not occur. Finally $P(0)$ is simply given as 1 less the sum of $P(n)$ for all $n \neq 0$ so that the total probability sums to 1.

At each step of the simulator, then, this distribution was sampled once for each phasotype with the sample size being equal to the current number of cells of that phasotype (the 'mutate' step in figure 6.1(b)). This is done for every phasotype in the initial population to produce potentially switched populations that started from each phasotype. These are then combined to give the "new" half of the population which is subject to switching. Adding this to the original population (the 'combine' step in figure 6.1(b)) produces a doubled population with the desired property that switching is applied to half of the new population with independent probability of changing at each loci.

A maximum population size was selected that approached the expected sizes for the experimental data sets whilst maintaining an acceptable speed of execution. Thus growth was allowed to proceed to a limiting size of approximately 10^9 cells

(i.e. a maximum of 30 generations with the exact size being dependent on whether the initial population size was a power of 2). Bottlenecks were applied by randomly selecting a number of individuals from the population with each cell in the population having an exactly equal chance of being carried through the bottleneck to the next generation. This was done by sampling from the discrete uniform distribution to select individual cells without replacement. This subpopulation was then used to seed the next generation and begin a new cycle of growth. This bottleneck could be repeated by allowing the population to grow to its maximum size before re-applying the bottleneck. The model was run multiple times using a constant input population. Phase switching and bottleneck selection in this model are fully stochastic so each run of the model produces a different output and thus the results presented in the rest of this chapter are from multiple runs of the model.

Unless otherwise stated, the input population was generated as representative of a colony seeded by a single cell. This was done by seeding the simulator with a single cell with phasotype `000000` and then growing this to the maximum population size (i.e. for 30 generations). Because mutation is modelled as having equal probability of changing phasotype in both directions, the choice of initial phasotype is arbitrary and does not affect the outcome. In order to generate a representative population this was repeated the same number of times as the following simulations (i.e. 50 or 100) and the diversity calculated for every repeat of this simulation. The population with the median diversity value among these repeats was chosen as representative and used as the input population.

6.3.3 Quantifying the changes in population

Because each *in silico* or *in vitro* experiment produced a population containing varying proportions of different six-gene phasotypes, it was necessary to summarise these differences so that the patterns of change could be understood. Changes in

the phasotypes of each population were quantified along two axes: (1) the number and proportions of different phasotypes within the population which is referred to as diversity; and (2) the difference between the input and output populations, which is referred to as divergence (see figure 6.2).

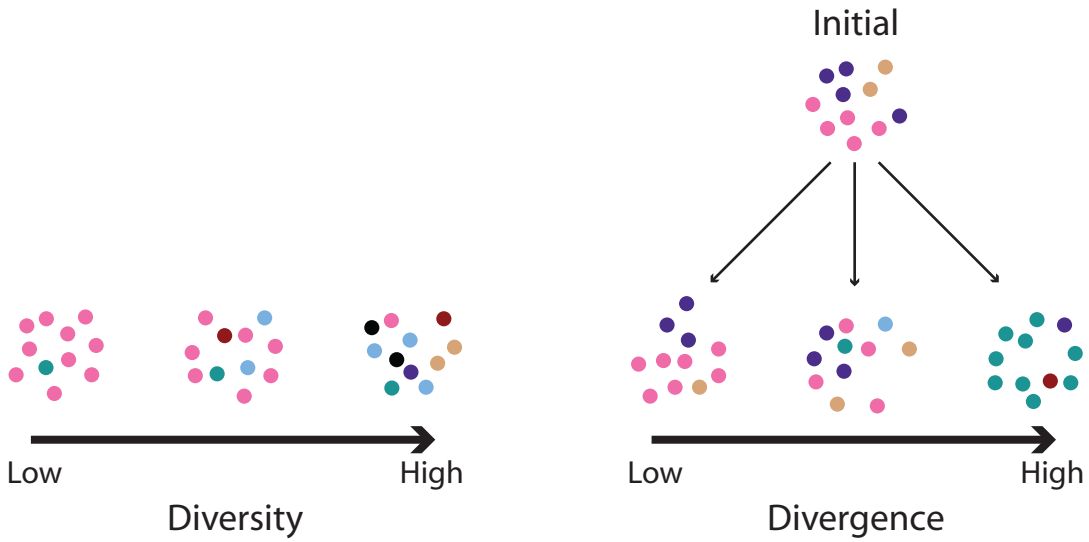

Figure 6.2: Illustration of the difference between diversity and divergence
Circles represent cells, with different colours corresponding to different phasotypes. On the left are three populations with increasing diversity going from left to right. The low diversity population is dominated by a single phasotype whilst the high diversity population contains many phasotypes in roughly equal proportions. This is a within population measure whereas divergence is a between population measure, thus the three populations shown on the right are increasingly diverged from the 'Initial' population above going from left to right. The first population has low divergence because it contains the same phasotypes as the initial population and in similar proportions, while the high divergence population shares no phasotypes with the initial population. Note that this high divergence population has low diversity since it is dominated by a single phasotype.

The 'diversity' was quantified using Shannon Equitability, which is simply Shannon Entropy (also known as the Shannon Index; Shannon 1948) normalized by its maximum possible for the dataset to give a number between 0 and 1 where 0 is the minimum possible diversity (the entire population has a single phasotype) and 1 is the maximum possible diversity (every possible phasotype is present in equal amounts).

The 'divergence' was quantified using population separation, which is simply the proportion of each population not shared between the two populations. This gives a number between 0 (the populations have exactly the same phasotypes present

in exactly the same proportions) and 1 (the populations have no phasotypes in common).

Mathematically, these are expressed as follows:

$$\text{diversity} = S/S_{\text{max}}$$

where:

$$S = -\sum \log_2 p_i^{p_i}$$

where p_i is the proportion of the population in each phasotype and S_{max} is the maximum value of S on this dataset. This is achieved when the phasotypes are present in equal quantities and so if the sample or population size, N, is less than the number of phasotypes (2^G where G is the number of genes) this is:

$$S_{\text{max}} = -\sum_{1}^{N} \tfrac{1}{N} \log_2 \tfrac{1}{N} = log_2 N$$

or, if the number of phasotypes is less than, or equal to, the sample or population size:

$$S_{\text{max}} = -\sum_{1}^{2^G} \tfrac{1}{2^G} \log_2 \tfrac{1}{2^G} = log_2 2^G = G$$

Divergence is most easily calculated by calculating the population overlap and then subtracting this from 1 to give population separation, this is calculated as:

$$\text{divergence} = 1 - \sum min(p_i, q_i)$$

where p_i and q_i are the phasotype proportions for each population.

6.3.4 Bottleneck size has qualitative and quantitative impacts on the output population

The effect of five repeated bottlenecks on virtual populations was modelled with six genes and a mutation rate of 1 in 500, which is within the estimated range of phase variation rates for *C. jejuni*. After each bottleneck the population was grown to a size of 2^{30} before the next bottleneck was applied. The simulator was run 100 times for each bottleneck size and then diversity and divergence scores were calculated for each of the output populations with the latter measuring differences from the initial population. The diversity versus divergence score of each output population was plotted in order to visualize the internal and temporal effects of different bottleneck sizes on population structure. Each point on figure 6.3 is the outcome of one run of the simulator. These data show that bottleneck size produces major quantitative and qualitative shifts in the output populations. Thus, small bottlenecks produced a bimodal pattern with distinct output populations of high and low divergence but both of low diversity. Contrastingly larger bottlenecks result in convergence on populations with intermediate levels of both divergence and diversity. Divergence is still generated with the larger bottlenecks but is now due to the diversity generated during replication from a population that becomes more diverse with every replication rather than from a direct effect of the bottleneck.

Bottleneck size also impacts population structure. With single cell bottlenecks and the larger bottlenecks the population structure is similar in all output populations whereas intermediate smaller bottleneck sizes produce disrupted population structures. This is illustrated in figure 6.4, where six example output populations derived from the same input population with 1, 8 or 1024 cell bottlenecks are shown. In the case of 1 and 1024 cell bottlenecks, the population contains a single dominant phasotype and then a cluster of phasotypes closely derived from this phasotype,

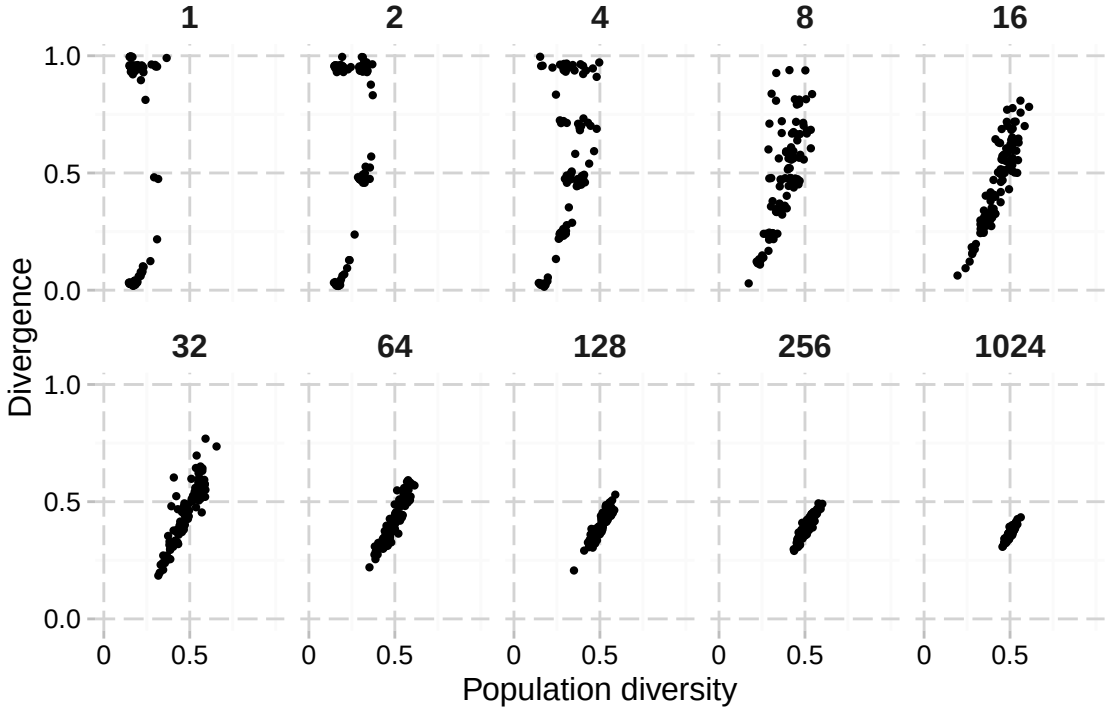

Figure 6.3: Divergence and diversity in simulated data with a range of bottleneck sizes

Each point on the figure represents an individual run of the simulator with 100 runs for each bottleneck size. The model simulates a six gene phasotype with each gene switching at 0.002 mutations per division for both directions of phase variation (i.e. ON-to-OFF and OFF-to-ON). The starting population has an initial diversity of 0.13. Each population was grown to a size of approximately 10^9 (30 divisions). Diversity is measured by Shannon Equitability, divergence is the difference from the initial inoculate quantified as population separation.

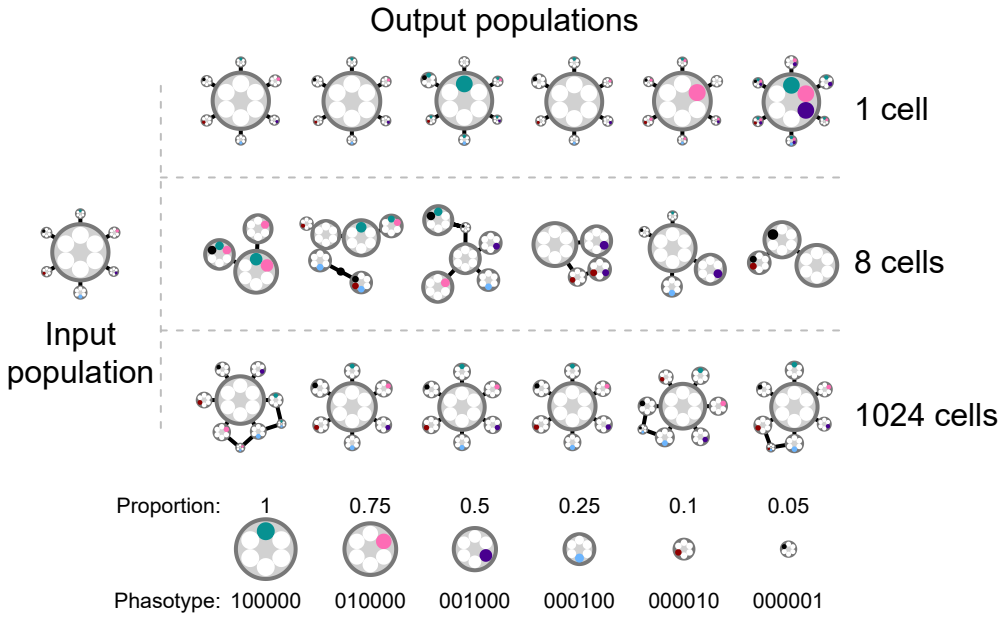

Figure 6.4: Changes in output population structure after application of bottlenecks
Figure shows the population structure from six independent runs of the simulator under three different bottleneck sizes. The common input population is shown on the left. Each population is represented as a group of linked circles, where each circle represents a different phasotype with the area of the circle proportional to the proportion of the population in that phasotype. The six inner circles represent the phasotype with filled (coloured) circles representing the ON state for that gene, and empty (white) circles representing the OFF state. All single gene changes are linked by solid lines, while those lines with a circle in their middle represent two gene changes and are only included where necessary to produce a connected graph. Phasotypes that compose less than <1% of the population are omitted. These omitted phasotypes make up, in total, an average of 0.6% of the population for 1 cell bottlenecks, 5.1% for 8 cell bottlenecks and 7.5% for the 1024 cell bottlenecks.

whereas in the 8 cell bottleneck the population structure is disrupted and contains multiple phasotypes which each contribute a large proportion of the population. In the case of the 1024 cell bottleneck the major phasotype is always the same major phasotype as that present in the input population; contrastingly roughly half of the output populations from the single cell bottleneck have a new major phasotype. Critically, this new major phasotype is not only novel with respect to the input population but differs between output populations. This new major phasotype also makes up a large proportion (>85%) of the population in these single cell conditions. Rare phasotypes (those where the individual phasotype composes < 1% of the population) also make up an increasing proportion of the total population as the bottleneck size increases (0.6% of the population for 1 cell bottlenecks, 5.1% for 8 cell bottlenecks and 7.5% for the 1024 cell bottlenecks).

In some cases, the significant factor may not be the diversity within the population but simply the presence or absence of particular phasotypes. To investigate the chances of a particular phasotype being generated, the coverage of the space of possible phasotypes was calculated with a range of bottleneck sizes and simulated generations. Table 6.1 shows the mean number of phasotypes generated and the proportion of 100 runs in which all phasotypes were present in the final population. As expected the coverage increases with bottleneck size but the impact of the final population is more profound. With a large final population (2^{30}), the mean coverage was 62.80 phasotypes even in the single cell bottleneck – or almost complete coverage of the 64 possible phasotypes with 6 genes. Note that the coverage is not uniform with the most distant phasotypes being less likely to be reached. Since in a symmetrical mutation model the starting phasotype is arbitrary, the starting phasotype is always `000000` and the most distant phasotype is always `111111`.

	Final population size					
	2^{24}		2^{27}		2^{30}	
Bottleneck	% All	mean	% All	mean	% All	mean
1	0	50.54	0	58.55	15	62.80
2	0	52.27	6	60.35	37	63.26
8	2	58.04	26	62.30	84	63.83
128	5	60.18	51	63.40	97	63.97
1024	11	62.24	87	63.87	100	64.00

Table 6.1: Coverage of the phasotype space changes with bottleneck size
Table shows the mean number of phasotypes present in the final population from 100 runs of the simulator, and the percentage of those runs in which all possible phasotypes (64) were produced in the final population. Because the simulator is fully stochastic repeating the sample of a hundred runs of the simulator will give a different percentage of runs resulting in total coverage and slightly different mean. The values given here are representative.

6.3.5 Varying the number of genes and mutation rate has a limited effect on the impact of bottleneck size

As there are significant variations in both the mutability and number of phase-variable genes within and between *C. jejuni* genomes, the model was run with a series of mutation rates ranging from 1 in 200 to 1 in 1500 mutations per division and between 3 and 10 genes. These parameters had only minor effects on the qualitative properties of the output populations (figures 6.5 and 6.6, respectively). Quantitative effects were observed with higher numbers of genes increasing the divergence produced by smaller bottlenecks. Similarly, and as expected, higher mutation rates increased both the diversity and divergence across all bottleneck sizes. Opposite effects were produced by smaller numbers of genes or lower mutation rates.

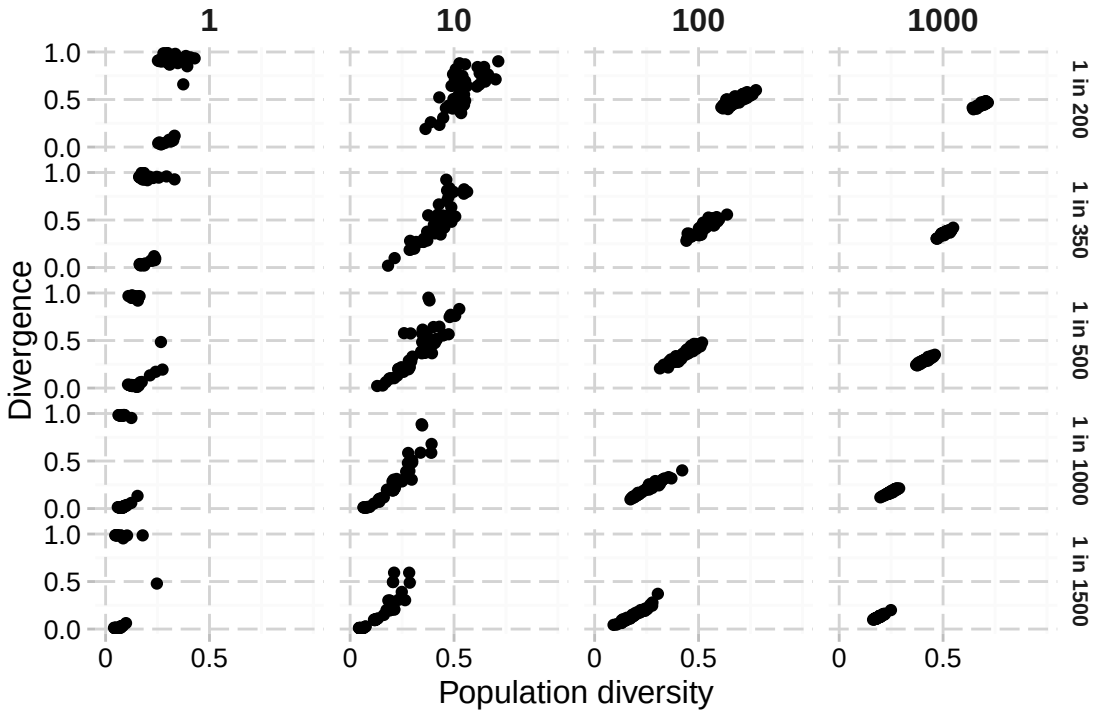

Figure 6.5: Impact of mutation rate on simulated populations
Each point represents a single run of the simulation which was run 50 times for each combination of bottleneck size and mutation rate. Bottleneck size is shown at the top increasing from left to right, while the mutation rate per division is shown on the right decreasing from top to bottom.

6.3.6 Changing the initial population structure impacts the effect of population bottlenecks

This first set of simulations were seeded with an initial population with a low level of diversity, created by generating a hundred populations from a single cell and choosing the one with the median diversity as the starting population. These simulations represented a typical population arising from a single cell and mimic experimental data sets. However, natural populations exhibit a wide range of diversities. To examine the impact of initial population diversity, the simulator was run with several different starting populations (figure 6.7). The output populations of the 1,000 cell bottleneck displayed decreasing divergence but increasing diversity as the diversity of the initial population was increased. A similar, but less pronounced, pattern is observed with the 100 cell bottleneck but with the 10 and 1 cell bottlenecks the more diverse starting populations only increased divergence

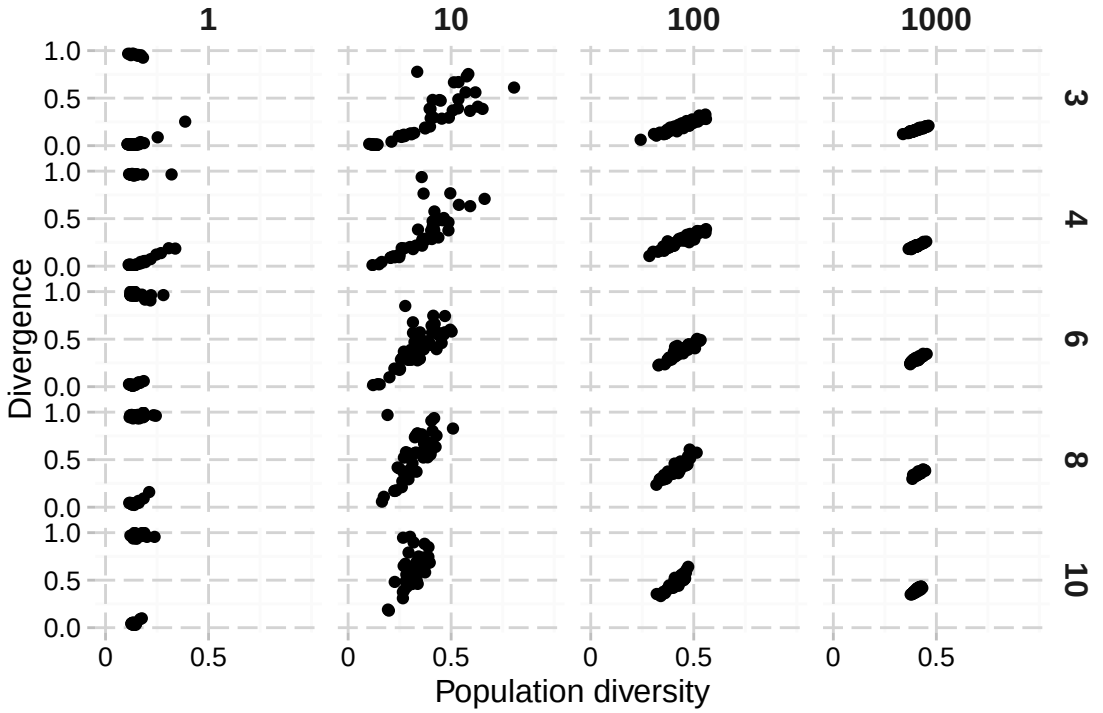

Figure 6.6: Impact of number of genes on simulated populations
Each point represents a single run of the simulation which was run 50 times for each combination of bottleneck size and number of genes. Bottleneck size is shown at the top increasing from left to right, while the number of genes is shown on the right increasing from top to bottom. Note that because the maximum possible number of phasotypes increases with the number of genes the diversity scores are not directly comparable between populations with different number of genes simulated.

but not diversity. This is due to these small bottlenecks restricting the amount of diversity that can pass from one passage to the next. Strikingly, with a maximally diverse population, the bimodal distribution seen with small bottlenecks disappeared and all populations exhibited the maximum divergence from the input population. There was also high divergence between output populations.

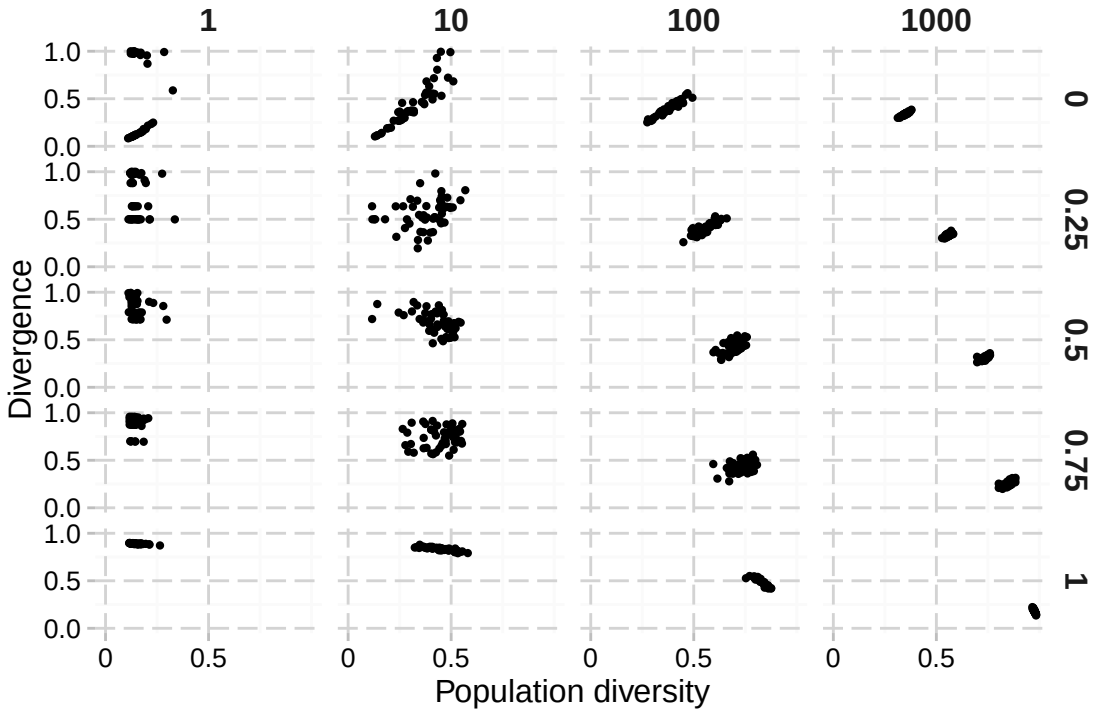

Figure 6.7: Impact of bottlenecks depends on initial population structure
Each point represents a single run of the simulation which was run 50 times for each combination of initial population diversity and bottleneck size. Other parameters were as set in figure 6.3. Bottleneck size is shown at the top increasing from left to right, while initial population diversity is shown on the right increasing from top to bottom.

6.3.7 Experimental testing of the impact of repeated bottlenecks

Note: The experimental work described here was carried out by Shweta Rajopadhye and Nwanekka M. Akinyemi before I joined the lab. I then re-analysed approximately half of all experimental samples to obtain complete phasotypes, and the data was re-analysed by Dr. Lea Lango-Scholey and I.

To test how non-selective bottlenecks influence the genetic structure of phase-

variable gene expression patterns, *C. jejuni* strain NCTC 11168 was subjected to multiple passages on MHA plates and interposed bottlenecks between each passage. A bottleneck involved random picking of one to 1,000 colonies, which should not impose selection on any of the phase-variable loci. Two experiments were performed, each of which started from separate inocula and involved application of five repeated bottlenecks to one to five separate lineages. A representative sample of 30-60 colonies was picked from each inoculum and output population. Repeat numbers were determined for 28 phase-variable loci and utilized for derivation of binary (0, 1) expression states. The expression states of the multiple colonies were converted into a %ON for each sample (figure 6.8).

The 1 cell bottleneck exhibited major shifts in %ON of some genes of the output populations as compared to the inoculum but with a random pattern for each lineage. For example, lineage NE2C had an ON-to-OFF switch in *cj1318* and *cj1426c* while NE5F has an OFF-to-ON switch in *cj1310* and an ON-to-OFF switch in *cj1421c*. Switches in larger bottlenecks tended to be partial changes in proportions with more homogeneity between lineages.

6.3.8 Comparison of experimental and *in silico* phasotype patterns

The total number of possible phasotypes for 28 phase variable genes is $2^{28}(\sim 3\times10^{8})$. However, the sample sizes collected (n=30) limit the detection of statistical significance to analysis of 32-128 phasotypes depending on population structure. Accordingly, 24 of the tracts were assigned to 4 groups of 6 genes for independent comparison to the *in silico* results (table 6.2). Four genes were excluded due to incomplete data sets for some samples. These groups were chosen to have approximately equal proportions of genes with different repeat numbers (since this factor is a major determinant of PV rate) and with a similar distribution throughout

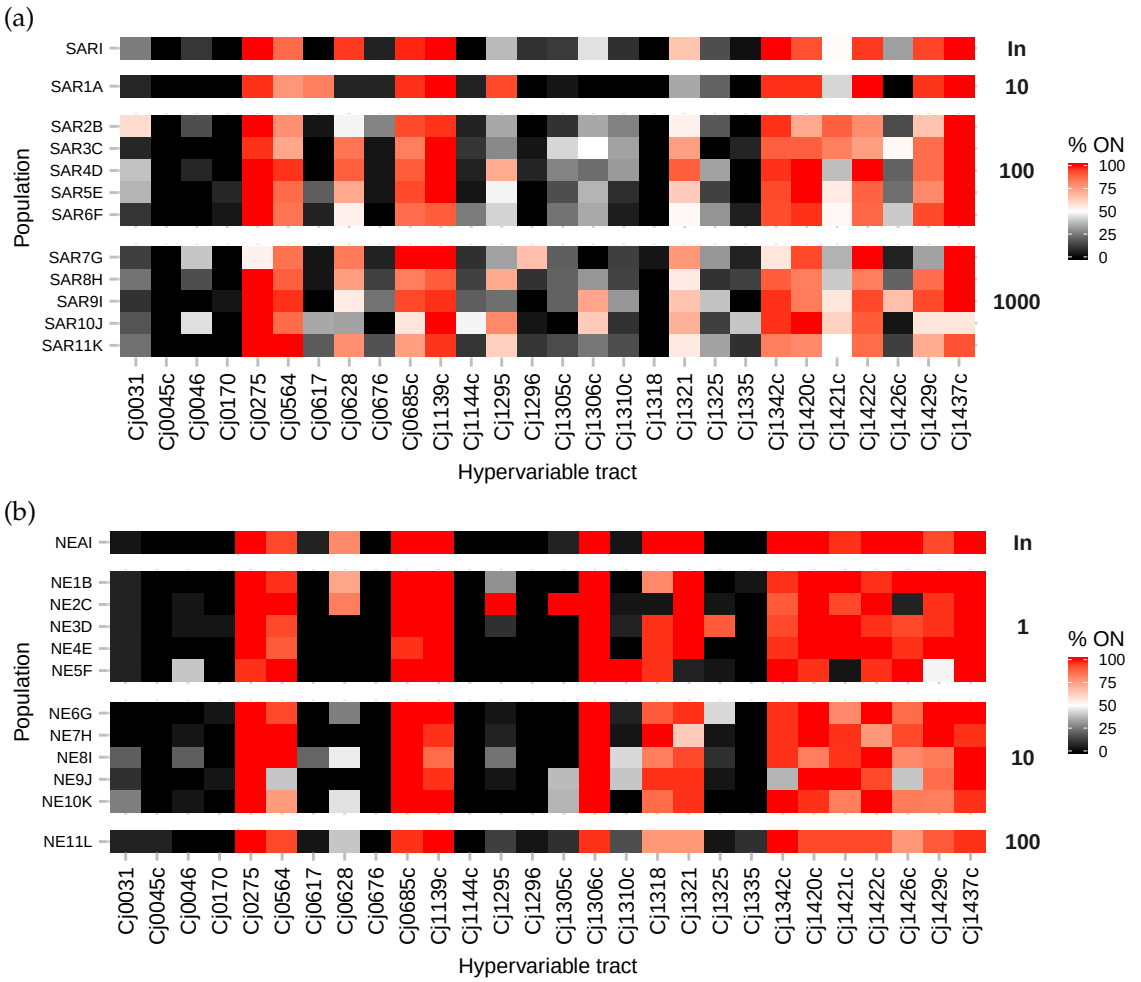

Figure 6.8: Changes in %ON during *in vitro* bottleneck experiments
Populations of *C. jejuni* strain NCTC 11168 were subjected to a series of five passages on Mueller Hinton agar plates. Bottlenecks were imposed between each passage by harvesting colonies from serial dilutions of each population. Samples of 30 colonies were collected for the inoculum and each lineage. For every phase-variable gene of each colony, expression states were determined from repeat numbers using the 28-locus PV assay. For each population, the proportion of each gene in the ON state was converted to a percentage and plotted on a color scale. Labels to the left indicate individual lineage, labels to the right indicate the bottleneck size imposed, or 'In' for the initial inoculum. Figures (a) and (b) are for two separate experiments in which all later populations are started with the same inoculum (SARI and NEAI, respectively).

the genome. These groups will behave independently if phase variation operates independently at each locus and the experiment was non-selective. Both of these assumptions are thought to be correct, as the experiment was designed to be non-selective, however if either assumption is violated then the behaviour of the experiment and model will diverge.

Group A	Group B	Group C	Group D
cj0676 (10)	*cj0045c* (11)	*cj0170* (8)	*cj0031* (9)
cj1144c (10)	*cj0617* (10)	*cj0564* (10)	*cj0275* (8)
cj1296 (10)	*cj1139c* (8)	*cj0685c* (9)	*cj0628* (11)
cj1310c (9)	*cj1305c* (9)	*cj1306c* (9)	*cj1295* (9)
cj1325 (9)	*cj1321* (10)	*cj1342c* (9)	*cj1318* (11)
cj1422c (9)	*cj1426c* (10)	*cj1429c* (10)	*cj1420c* (9)

Table 6.2: Genic composition of *C. jejuni* phasotype groups
Composition of the four phasotype groups used to compare experimental and *in silico* results. Numbers in brackets after gene names are the ON lengths for each gene.

In order to make the in *silico* model mimic the experimental model, the initial population for the simulated runs was seeded with the average proportions from the inocula used in the experimental data and then grown to the maximum population size before the first bottleneck was applied. A mutation rate of 1 in 300 was chosen as approximating the expected average mutation rate of the actual loci.

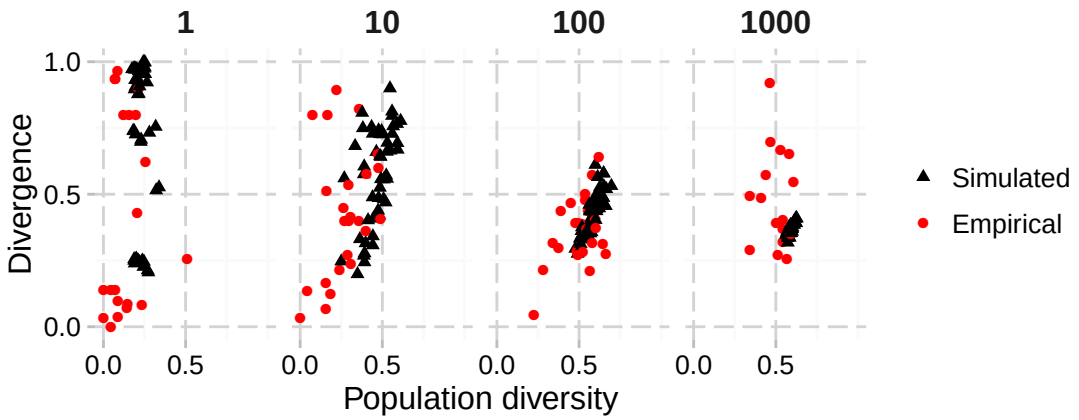

Figure 6.9: Comparison of *in silico* and *in vitro* results
Each black triangle represents a separate run of the simulator, while each red dot shows the results generated by grouping genes from the empirical populations calculated from 16-30 picked colonies. The numbers along the top indicate the size of bottleneck applied. One lineage, SAR11K (1000 cell bottleneck), was omitted due to poor data quality.

Divergence and Diversity figures were calculated separately for each sub-phasotype for the experimental data and then compared to the output of the model (figure 6.9). The data for the small bottlenecks was similar to the model outputs but increasingly dissimilar for the larger bottlenecks. The experimental data for the 10 cell bottleneck exhibited lower than expected diversity for most of the populations while 40% of the output populations of the 1,000 cell bottleneck had higher than expected divergence. In order to determine whether these differences might result from a systematic effect of sampling from the populations, the *in silico* model was run with random sampling from the output population and the diversity and divergence values obtained from these samples compared to the true population diversity and divergence values. No systematic bias was detected for the sample sizes used in the *in vivo* experiments (data not shown).

6.4 Modelling of cyclical selection

Having explored non-selective bottleneck selection, I now move on to discuss modelling aimed at mirroring the cyclical selection assay. Many of the same principles underpin this model but with important changes to achieve three new goals: (1) the modelling of differences in mutation rate due to tract length changes, (2) the introduction of selection based on the ON/OFF state of the individuals in the population, and (3) to represent a population growing to a limit rather than a continuously doubling population.

6.4.1 Changes in model to represent alternating selection

Because the cyclical assay operates only on a single gene, only a single gene needs to be modelled, but this gene is now modelled as having a tract length of between

7 and 13 nucleotides as well as separate, non-PV, always ON and always OFF states (in some simulations). The mutation rate for each state is modelled with a distinct probability of increase or decrease in tract length based on its current tract length and empirical data from Bayliss *et al.* (2012). Unfortunately this dataset does not cover every possible eventuality so some values were approximated (see table 6.3) based on the available data. The lengths were capped at 7 above and 13 below by setting the probability of extending beyond these ranges to zero. Tracts larger than 13 are vanishingly rare ($< 0.5\%$ of all poly-G/C tracts according to the analysis described in chapter 3), whilst tracts lower than this range are not usually considered phase variable. The mutations were then applied to counts of each tract length by picking random numbers from the binomial distribution, or the normal approximation to the binomial distribution for large numbers, to calculate the number of switchers in the upward and downward direction separately.

Tract length	Insertion rate	Deletion rate
7[1]	0.0001	0.0
8	0.0004	0.000 04
9	0.001	0.0002
10	0.0004	0.0018
11	0.0002	0.0028
12[2]	0.0001	0.003
13[1 3]	0.0	0.004

[1] - Rate set to 0 to clamp possible range of tract lengths to 7-13
[2] - No experimental data available for increase, rate chosen to approximate trends
[3] - No experimental data available for decrease, rate chosen to approximate trends

Table 6.3: Simulated rates of increase and decrease
Rates of increase and decrease per generation for different tract lengths. The rates in this table are derived from Bayliss *et al.* (2012), tables 1 and 2. Note that Bayliss *et al.* (2012) gives broad ranges for these rates but the central estimates have been used here, and where more than one estimate for the same length has been generated these have been combined.

Selection is applied to each cell as a chance for that cell to survive in either the ON or OFF state which depends on whether the trait is undergoing positive or negative selection. This is parametrised as selection strength, s, where the probability of a cell surviving is $1 - s$ and thus a higher s indicates stronger selection. This differs from the parametrisation of models such as that in Palmer *et al.* (2013) in that there

the parameter s is used to directly multiply the size of the population followed by a normalisation to the maximum population size. However this difference is not functionally distinct except where the selection results in the population reducing below the maximum possible size. The number of cells of a certain length surviving can again be calculated by picking numbers from the binomial distribution (or the normal approximation to it) ensuring that the simulation can be rapidly performed. For most of the simulations discussed below the selection is modelled with all cells of the selected-for state surviving and a proportion of the cells in the selected-against state dying. Selection in one direction is applied for a number of generations, before the conditions are switched to select in the other direction, and by repeating this alternation the selective cycle is simulated.

In order to impose a limit on population growth, a survival chance is applied to each cell before division occurs. This chance depends on the total size of the population, not on the number of cells of a given length nor the state of the cells, and is defined to be 50% at the population size limit and 100% for a single cell and vary linearly between these points. Since the population doubles at division, this holds the population approximately steady at the population limit although small fluctuations occur because the chance of survival is applied stochastically, again using random numbers chosen from the binomial distribution. The limit can be varied, but was chosen as 10^{10} for all simulations discussed below.

6.4.2 Validation of model by stationary point

Because the rate of flow between the populations of different tract lengths depends linearly on the size of those populations there will exist an equilibrium population such that these flows cancel each other out. In the absence of selection, the population is expected to eventually reach this point, and thereafter mean changes in the size of the population of each tract lengths will even out and only minor stochastic

variation in the population population will occur. This equilibrium population is the stationary point of the model.

The stationary point can be analytically determined by creating a matrix from the expected values of the change probabilities, X and then finding a population, expressed as a vector p such that $Xp = p$. This will have a range of values but if the population values are treated as a proportion of the whole, rather than counts of individual cells, then we have $\sum p_i = 1$. The solution can then be found by finding the eigenvector of X with an eigenvalue of 1 and normalising the resulting vector. Note that this stationary point is independent of the starting population so it can be compared to the stationary distribution of the model simply by leaving the model running until the population levels approximately stabilise from any starting point. This serves as validation of the model. Figure 6.10 shows the behaviour of the model under conditions of no selection and indicates that the proportions of each tract length asymptotically approach the stationary point. Continuation of the model allows the stationary point to be determined from the end state of the simulator after 500,000 generations. These proportions can then be compared to the numbers analytically derived and, as indicated in table 6.4, these values show complete correspondence between the two methods confirming that the model is behaving in the expected fashion under the condition of no selection.

A striking feature of this stationary point is the high proportion of the G9 tract length that it contains (48.8%) with a further 41.1% in the two adjacent populations (G8/G10). The longest tract lengths (G12 and G13) are almost completely absent from the stationary point. Neutral drift towards these favoured tract lengths, therefore, is likely to have a significant impact on the behaviour of the model.

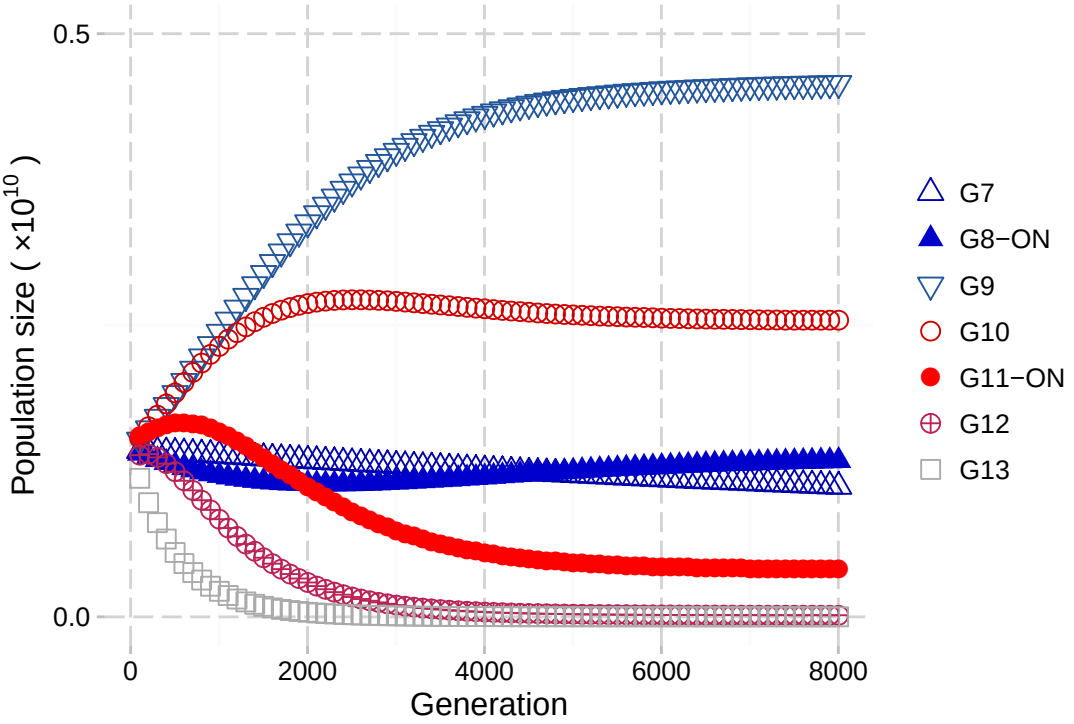

Figure 6.10: Population tends to a stationary distribution over time
Figure shows a single run of the simulator for 8000 generations without selection (points shown every 100 generations). The stationary distribution is derived from the final population numbers; changing the initial population does not alter this stationary outcome although it may alter the time to reach it (data not shown). Mutation rates for each tract length are as shown in table 6.3. Filled shapes indicate ON states but since no selection was applied this distinction is arbitrary. Not that the 8000 generations shown is insufficient for the population to reach the final stationary point and, in particularly, the G7 proportion will drop from the shown point to a lower level.

Tract length	Analytic	*in silico*
7	0.056	0.056
8	0.140	0.140
9	0.488	0.488
10	0.271	0.271
11	0.043	0.043
12	0.001	0.001
13	0.000	0.000

Table 6.4: Analytical and *in silico* stationary populations
Table shows proportion of stationary population in each tract length, comparing the analytically derived stationary population, and stationary population derived from final population of simulator run without selection for 500,000 generations starting from a uniform population. Results are shown to 3 decimal places. Simulated population shows continual random variation around the stationary population and convergence of the two extreme populations (length 7 and 13) to the analytic prediction (accurate to 3 decimal places) takes over four hundred thousand generations.

6.4.3 Searching the parameter space under symmetrical conditions

Alternating selection can be either symmetrical or asymmetrical, depending on whether the selection strength and time are the same under both selective conditions. As the simpler case, symmetric conditions were the first to be investigated. Here the simulation switches between favouring the ON and OFF states every T generations, and the selection against the disfavoured state, s, is constant throughout. Because the mutation rate depends on tract length, the choice of ON lengths is important. The results of two simulation runs are shown in figure 6.11 which illustrate the important features of the model. Here s was set to 0.05, 8G is ON and the model switches between selective states every 2000 or every 500 generations. The model begins by selecting for the OFF state, and the population rapidly comes to be dominated by lengths in the OFF-state (shown by the empty shapes) but over the period of stability, the ratio of the slower-varying (blue) lengths to the faster-varying (red) lengths approaches a constant defined by their relative mutation rates (as discussed above). After a period of time the model switches to selecting for the ON-state and the ON lengths (indicated by filled shapes) rapidly take over the population. In the two example runs, the slower varying environment (switching every 2000 generations) favours the shorter, slower varying tract whilst the faster varying environment (switching every 500 generations) favours the longer, faster varying tract.

These two examples were found by manually varying the parameters but a more systematic approach was needed to explore the general behaviour of the model. To do this the simulation was run for 16 cycles with a range of selection strengths and plotting the composition of the population after the final round of selection for the ON state. In each case, the initial population was generated by seeding

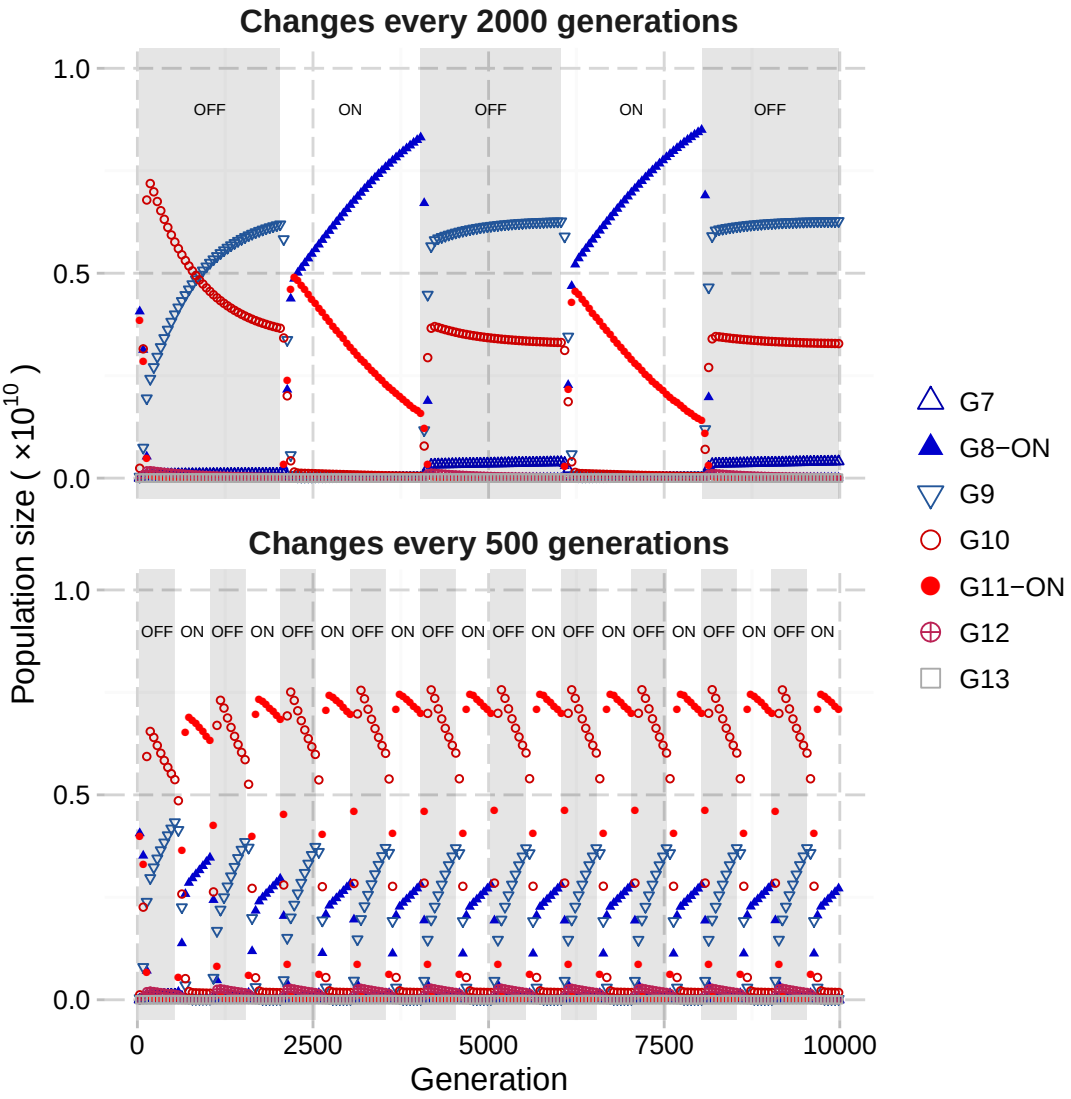

Figure 6.11: Favoured tract length depends on selective conditions
Two illustrative sample runs, showing changes in population structure under alternating selection for the ON and OFF states. In the top run the selection switches every 2000 generations, favouring the shorter (blue) tracts, whereas in the bottom run the selection switches every 500 generations, favouring the longer (red) tracts. Periods of ON selection are shown on a white background, periods of OFF selection on a grey background. Initial population contained the ON state followed by 32 generations of growth under no selection. In both runs, s is set 0.05 and G8 and G11 are ON states and are shown with filled shapes.

the population with 1 cell of each ON state and 1 non-PV locked-ON[1] cell and then allowing the population to grow without selection for 32 generations before applying the first round of selection. Any change in the final population from this equal distribution of cell types after the final round of ON selection would thus represent a fitness advantage for the favoured length.

The results of this parameter space search are shown in figure 6.12 for each of the three possible ON states. These will be referred to as G8-ON (with G8 and G11 being ON states), G9-ON (with G9 and G12 ON), and G10-ON (with G7, G10 and G13 being ON states). With all three ON state conditions, under low levels of selection, or rapid switching of selective conditions, there is little change in the population structure and the locked-ON cell type persists in the population. With high levels of selection this represents not an equilibrium but a steadily declining population size in which the population is unable to adapt to the rapidly changing conditions and were the simulation left to run long enough the population would reduce to zero. This happens in the top left of each figure, where the missing blocks indicate conditions under which the population is entirely wiped out.

There is a non-linear relationship between the length of the tract and its chance to increase or decrease in length. As a result there are qualitative changes in the behaviour of the model with different chosen ON lengths. With the G8-ON and G10-ON models which of the ON lengths is favoured varies with the strength of selection and period of stability. In the G8-ON model with low s or T, PV is not favoured over the non-PV variant (shown in green). With mild selection, once PV is favoured there is a period of increasing T in which the shorter tract length (G8-ON, shown in blue) is favoured before a period in which the longer tract length (G11-ON, shown in red) is favoured before at the highest levels of stability the shorter tract length once again dominates the population. Under higher levels of selection this

[1]Because the simulation here was symmetrical there was no need to include the locked-OFF state since its behaviour would simply mirror that of the locked-ON state.

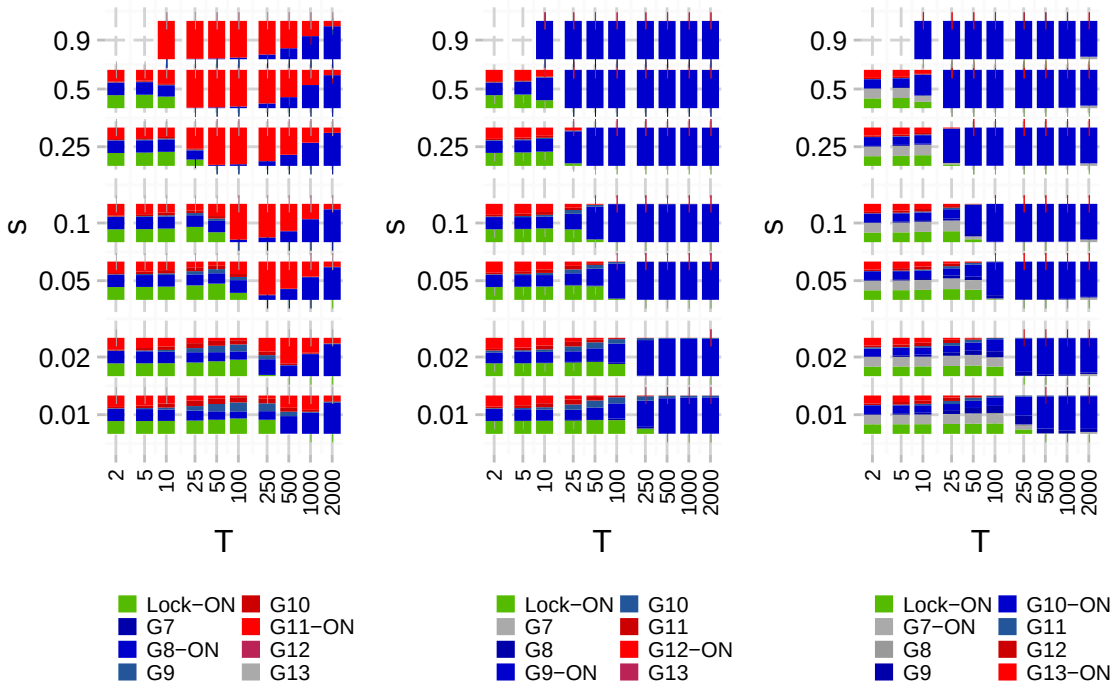

Figure 6.12: Response of output population to changes in length and strength of selection

Figure show the impact of changes in the s and T parameters on the resulting population under the three possible ON state configurations: G8-ON, G9-ON, and G10-ON (from left to right, respectively). In each case the figure shows the composition of the final population after 16 cycles of selection with the last selective cycle being for the ON state. Each bar shows the proportion of the population in each state. Colouration is different for each ON state, but in each case blue represents slower variation and red faster with dark green representing the locked-ON state; the neighbouring tract lengths to each ON length are given similar colours. For G10-ON, the third G7-ON state is shown in grey.

initial set of conditions that favour the shorter tract disappears and the proportion in the shorter tract length increases monotonically with increasing T. This pattern likely results from an interplay of neutral and selective forces.

In the G10-ON model, the intermediate G10-ON state (shown in blue) is favoured until extreme periods of stability are achieved (> 3000 generations, not shown) when the G7-ON length (shown in grey) starts to dominate the population. No set of conditions favours the G13-ON (red). Similarly, in the G9-ON model all conditions that favour PV result in the G9-ON length (shown in blue) becoming dominant over the faster switching G12-ON length (shown in red).

6.4.4 Impact of mutability estimates on simulated outcomes

Bayliss *et al*. (2012) gives a range of estimated mutation rates and uses four different reporter constructs or methods to derive these values. In most cases, there is approximate agreement between the estimates from different sources, however for G8 and G11 tracts this is not the case. The G8 estimates from the *cj1139-lacZ-cat* and *cj1139-lacZ-kan* constructs have non-overlapping 95% confidence intervals, as do the G11 estimates from the *cj1139-lacZ-cat* and the *capA* antibody assay. This section explores the impact of these different estimates on the behaviour of the simulation.

Estimates of G11 mutability materially alter simulated outcomes

The empirical dataset in Bayliss *et al*. (2012) contains two estimates for the mutability of G11 tracts, one based on a *cj1139::lacZ* reporter construct, the other based on a capA antibody assay. These two methods give highly divergent results for the mutability of the G11 tract. The former gives a rate of 40.45×10^{-4}, whilst the later

gives a rate two-and-a-half times lower at 16.41×10^{-4}, and the 95% confidence intervals for these estimates do not overlap. Moreover, the two assays also give divergent ratios of insertion to deletion, 23:1 for the lacZ reporter, and 5:3 for the antibody method.

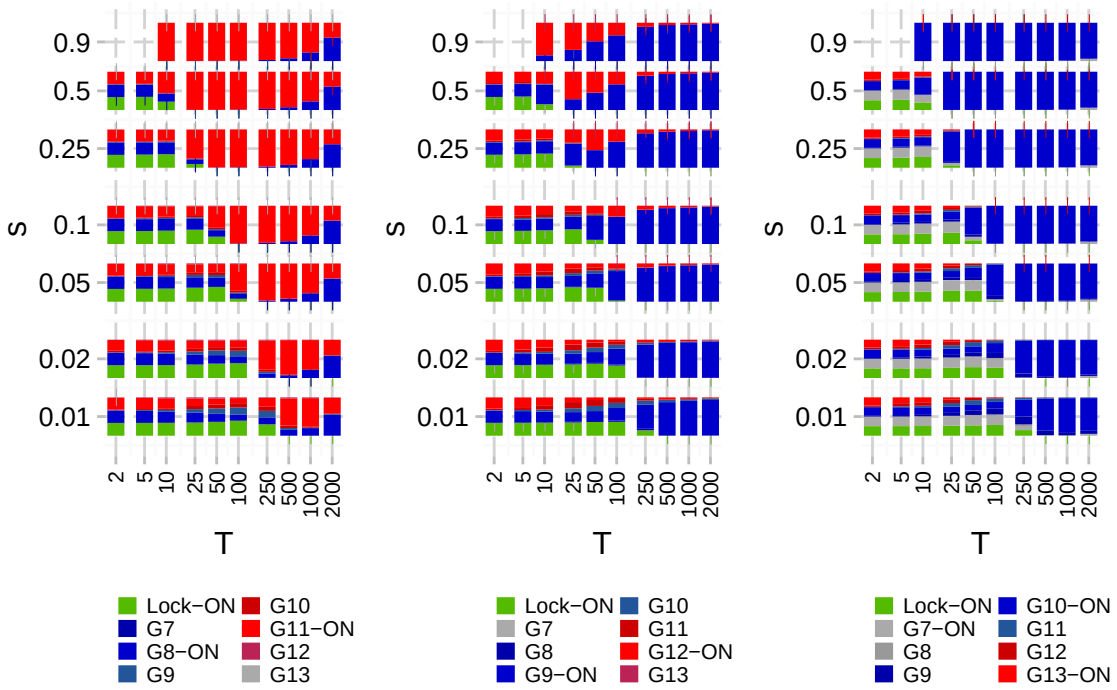

Figure 6.13: Parameter search space with *capA* derived mutability for G11
Figure shows the same set of parameter values as figure 6.12 but with the mutability of the G11 altered to reflect the mutability estimate from Bayliss *et al.* (2012) based on immunoblotting of the *capA* gene. Note the longer G12-ON tract is now favoured in the G9-ON simulation when the selection is strong and the period of stability short whereas in figure 6.12 there are no parameter pairings which favour the G12 tract in the G9-ON simulation.

The central model uses a value derived from the mean of these two methods but if, instead, the values for the *capA* gene are used the results markedly change (compare figure 6.13 and figure 6.12) with the more variable tract becoming much more competitive in the G8-ON and G9-ON simulations while the G10-ON model is largely unaffected by the change.

Estimates of G8 mutability do not materially alter simulated outcomes

In contrast, altering the G8 mutability to either the higher *cj1139-lacZ-kan* or lower *cj1139-lacZ-cat* estimate has little effect on the resulting simulations. These changes only result in mild increases or decreases in the proportion of any particular tract length observed under some conditions and do not result in any changes in favoured tract length (data not shown).

Uniform scaling of mutability has limited impact on simulated outcomes

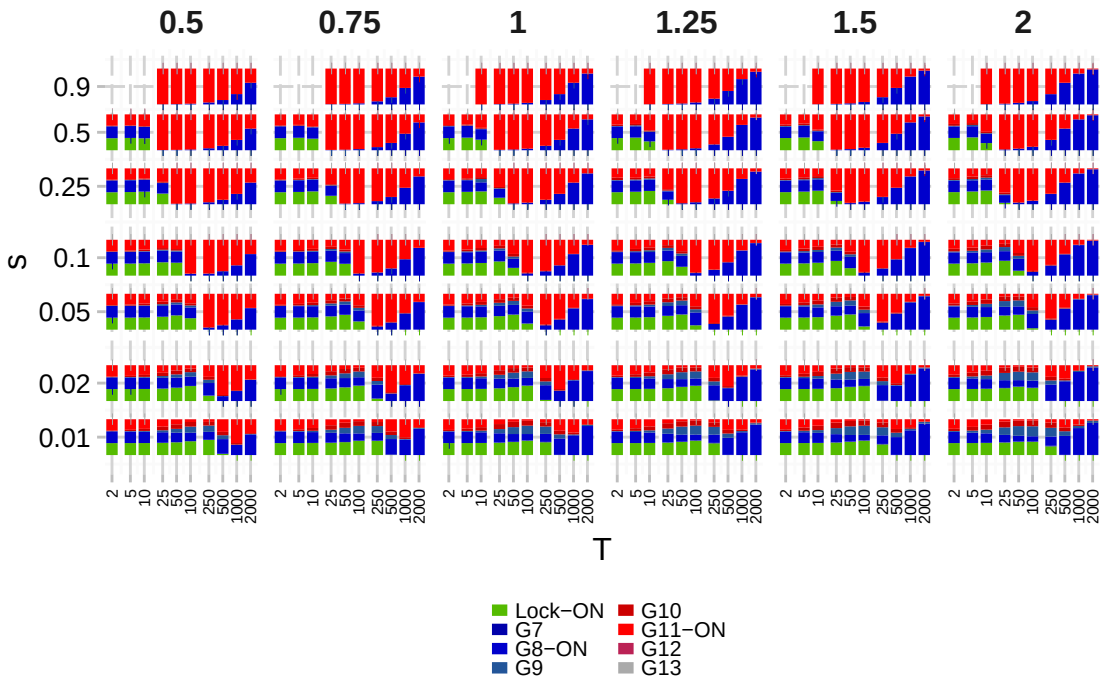

Figure 6.14: Effect of uniform scaling on G8-ON simulation
Figure shows the impact of uniform scaling of the mutability parameters on the G8-ON simulation. Values at the top show the scaling factor from 0.5 to 2.0. Other features as per figure 6.12.

In addition to the central estimate of mutability, Bayliss *et al.* (2012) also gives a 95% confidence interval for the range of the mutation frequency. Although the exact ratio of the 95% confidence interval range to the central estimate is not constant, the ratio of the interval ranges are completely contained within the range $\times 0.5$ to

$\times 2.0$. To explore the potential impact of these range of estimates, the simulator was run with values of the central estimate (table 6.3) uniformly scaled by values in the range $\times 0.5$ to $\times 2.0$. For the G8-ON and G9-ON models there was little impact across the simulated range of impacts, merely a very mild shift towards the shorter tract length as the scaling factor increased (see figure 6.14 for G8-ON, G9-ON not shown). For the G10 ON simulation, the highest scaling multiple brought the range favouring the shortest G7 tract down into the range of T values shown with the highest uniform scaling values (see figure 6.15).

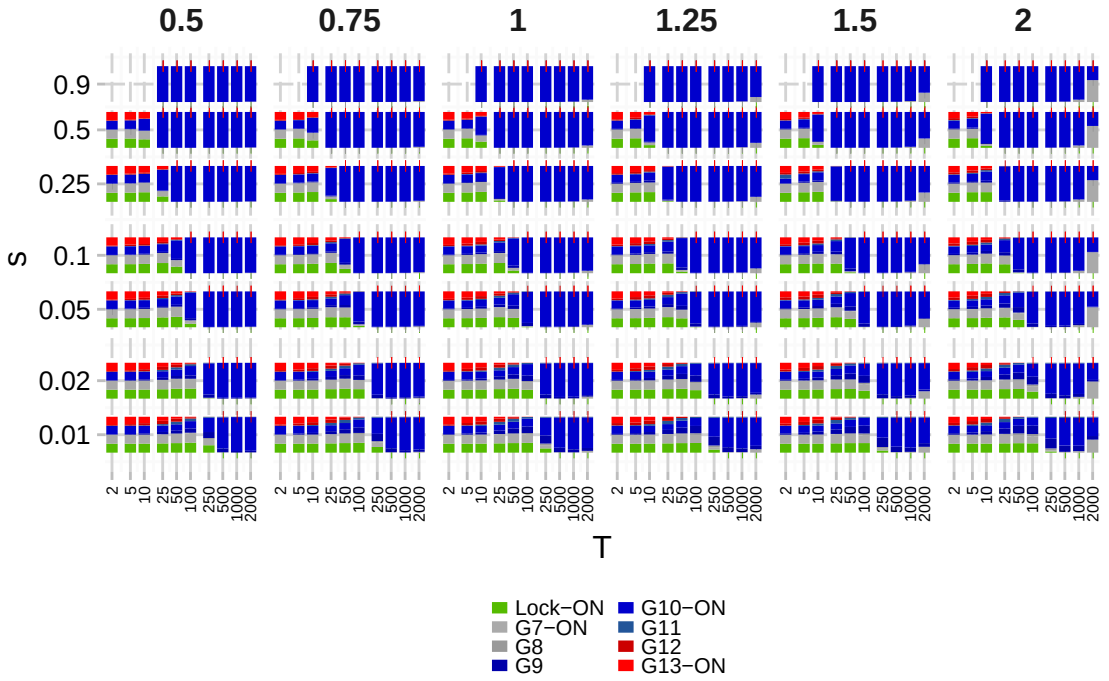

Figure 6.15: Effect of uniform scaling on G10-ON simulation
Figure shows the impact of uniform scaling of the mutability parameters on the G10-ON simulation. Values at the top show the scaling factor from 0.5 to 2.0. Other features as per figure 6.12. Note that the G7-ON tract length is favoured over the G10-ON at high T with the higher scaling factors.

6.4.5 Transitions between favoured ON length are largely determined by period of stability

The transition point between favouring phase variation and not favouring phase variation is determined by a combination of the s and T parameters but the transition between favoured ON lengths appears to be largely independent of the s value but rather depends almost solely on the T value. This is true for the G11-G8 transition and G10-G7 transitions which occur at epoch lengths, T of 2000 and 5000 generations (respectively), however the G12-G9 transition[2] occurs at an s-dependent point but the relationship between s and critical T appears non-linear.

6.4.6 Comparison of simulated tract lengths to observed tract lengths

These simulated transitions can be compared to observed tract lengths, both in *C. jejuni* NCTC 11168 (table 1.1) and more generally to the data gathered by Phasome*It* (figure 3.4, reproduced in figure 6.16 below for convenience). The tract lengths noted in table 1.1 are the nearest ON lengths, and contain a preponderance of length 9 tracts with a smaller number of length 8 and 10 tracts and a single G11 tract, broadly following patterns observed in the simulations. The close correspondence between the observed tract lengths in figure 6.16 and the distribution predicted to be the stationary point of neutral change in table 6.4 is even more remarkable. This suggests that averaged over many tracts and genomes, simple drift is the principal driving force behind tract length distribution.

[2]This transition only occurs with the G11-CapA mutation frequency, as discussed above, and it is this transition that is referred to here.

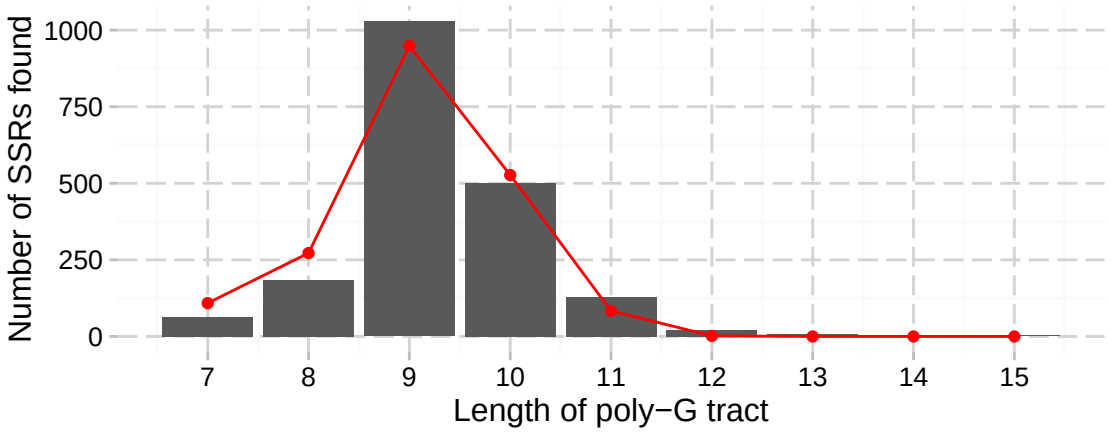

Figure 6.16: Distribution of poly-G tract lengths in *Campylobacter* genomes
Figure shows counts of poly-G tracts of each length found in the analysed *Campylobacter* genomes, from a total of 1944 tracts (see section section 3.5.1). Red line shows predicted distribution without selection based on empirical data of mutation rate for each tract length. Poly-G tracts of length below 7 were excluded from the search. Tracts that are poly-C in the direction of coding are also included. Figure reproduced from figure 3.4 for convenience.

6.4.7 Asymmetric selection conditions

Thus far, this section has concentrated on the behaviour of the model under symmetrical conditions where each form of selection is applied for an equal period of time and at the same strength. An alternative environment applies different strength of selection for OFF (s_0 and T_0) and ON (s_1 and T_1) conditions. Exploring this parameter space revealed that under conditions where the selection was strong in the OFF condition and weak in the ON condition, phase variable OFF states persisted through the ON condition. To investigate whether these conditions would favour a non-PV OFF state a second locked state was added to the model, this time simulating a locked-OFF state. With this second locked state in place the simulation now shows that these conditions that carried through the OFF states in fact favour a locked-OFF variant and thus conditions either favour PV or not (figure 6.17).

These asymmetrical selection results mirror those of Palmer *et al.* (2013) and support the conclusion that under asymmetrical conditions the selective effect must be sufficient to produce a complete sweep of the population in order for phase

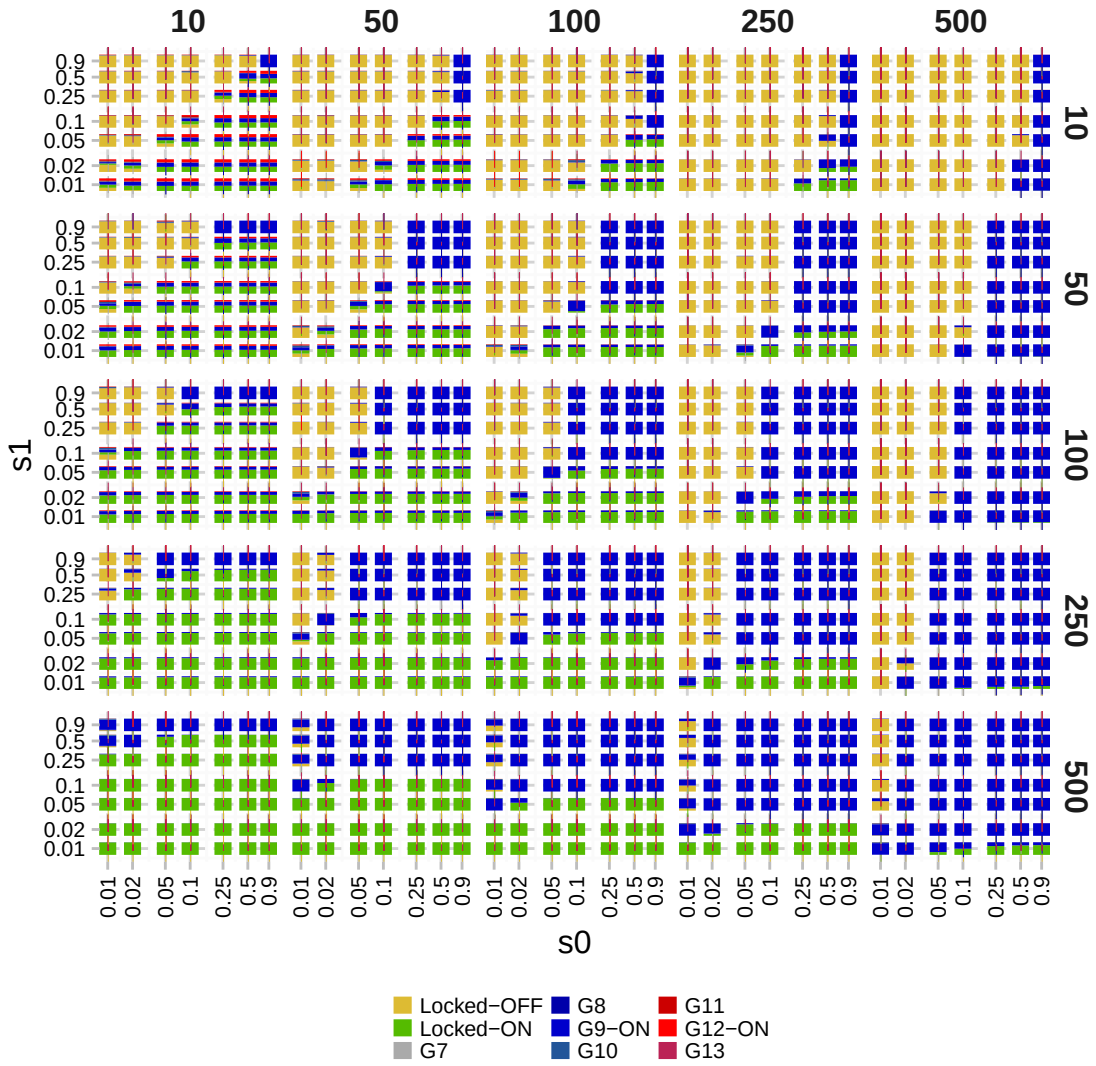

Figure 6.17: Impact of asymmetric s and T on G9-ON simulation
Figure shows impact of asymmetrical selection conditions. s_0 and s_1 values are the strength of selection against ON and OFF (respectively) when the environment is selecting for the opposite condition. The periods of stability are also asymmetric, T_0 increases from left to right, while T_1 increases from top to bottom. Each bar shows the proportion of the population in each state after the final round of ON selection following 16 cycles of alternating OFF/ON selection applied to an initial population containing equal amounts of each state. Note: only the two locked states and the G9-ON state are visible in the figure.

variation to be favoured but that this condition is relaxed in the case of symmetrical or near symmetrical environmental switching.

6.5 Simulation of *in vitro* cyclical selection assay

The proceeding section has given a broad survey of the behaviour of the cyclical model under a range of conditions. In this section I move on to discuss the application of the model to the cyclical assay developed in chapters 4 and 5. To employ the model, the first step is to derive experimental values for s under both conditions and then simulate a range of experimentally plausible variations in the conditions. These simulates can then be used to predict the outcome of the experimental assay.

6.5.1 Derivation of experimentally derived parameters

The parameters for the model can be analytically derived by calculating the degree of selection (s) that would produce the observed shift in proportion ON in the estimated number of generations occurring (T) in the experiment. For the phage selection this simplified approach means solving the equation $p_0(1-s)^g = p_1$ for s, where s is the selective force, g is the number of generations, and p_0 and p_1 are the proportion in the ON state before and after selection. By substituting $s' = 1 - s$ and re-arranging we get $s' = (\frac{p_1}{p_0})^{1/g}$. The experiment described in section 4.5 was run for 24 hours. The doubling time of *C. jejuni* is assumed to be approximately 90-120 minutes so 24 hours corresponds to 12-16 generations. For the case of complete selection ($p_0 \approx 0.98$ and $p_1 \approx 0.01$) this gives a range for s from 0.25 to 0.32. This range could be extended in both directions to account for the uncertainty inherent in the measurement of the proportion ON using the Clopper-Pearson mid-P interval to give 95% confidence intervals. However since the model requires

a single parameter value, a rounded value of 0.28 was used as the mean value in the range. In the experimental selective assay, the level of selection is likely non-constant during the phage selection step as there will be multiple rounds of phage infection resulting in the resultant MOI. Thus selection may be initially weak before increasing as the PFU/ml increases to, presumably, a maximum value. Applying a single constant value is thus a simplifying assumption of the model.

The serum selection step differs from the phage selection step as killing is effectively instantaneous selection and doubling of the population probably does not occur during this step. This step is treated inside the model as occurring during a single generation. It is also the case that the selected for ON state experiences a significant drop in number of cells, so rather than simply having a single selection co-efficient s there are two coefficients s_{on} and s_{off} with $s_{on} \ll s_{off}$. The ratio of these two coefficients can be determined from the ratio of ON to OFF before and after selection, while the exact values are based on the resulting CFU/ml. Using the information from the experiments in section 4.6, and again simplifying to single values this gives $s_{off} = 0.00004$ and $s_{on} = 0.02$.

6.5.2 Potential for constructing alternative ON lengths

These experiments will call for different tract length variants. These can easily be created either by inserting or deleting Gs from the homonucleotide tract in *cj1421c* either by mutagenesis or, alternatively, by simply picking mutant colonies of the desired length. In either case, it will be important to create multiple isolates and select one in which there has been no change in the other PV tracts. These different length isolates would still have G9 and G12 as the ON states; more interesting experiments will rely on the creation of alternative ON length constructs. With G8-ON (and thus G11-ON also) this can be done by inserting a single A, T or C at the end of the homonucleotide tract. This will change the ON frame without

altering the resulting DNA sequence since all triplets of the form GGN produce a Glycine residue. Unfortunately there is no sequence modification that would result in a G10-ON (and thus G7- and G13-ON) without altering the DNA sequence. However since the amino acid sequence tolerates changes in tract length without influencing function minor changes around this site may be tolerated. A deletion of the two nucleotides immediately following the site would remove a single Tyrosine residue and produce a G10-ON mutant. Interestingly, every homologue of *cj1421c* (as identified by the program described in chapter 3) has a predicted ON state of G9.

6.5.3 Experimentally plausible conditions will favour a phase variable SSR over a locked-ON tract

The range of possible conditions for empirical investigation are bound primarily by time, as the possibility for contamination increases as a function of the number of steps. To reflect these limitations the number of cycles has been reduced to five and the range of phage and serum selection steps chosen to those fit within a reasonable time period. Due to the fact that sera kills the ON state at high levels, a selection free recovery step is included after each period of serum selection to allow for some degree of population recovery before imposing the phage selection.

The results of these empirical simulations are shown in figure 6.18. It is immediately obvious that across most of the range of empirically plausible values the phase variable variants will not be favoured over the locked variants. However with longer periods of phage selection (2-4 days, i.e. 32-48 generations) and 2 or more cycles of serum selection the phase variable variants are predicted to be favoured. These experiments should be possible in practical terms but would require subculturing into new media during the phage selection step in order to maintain the growing population during phage selection. Unfortunately, these results also

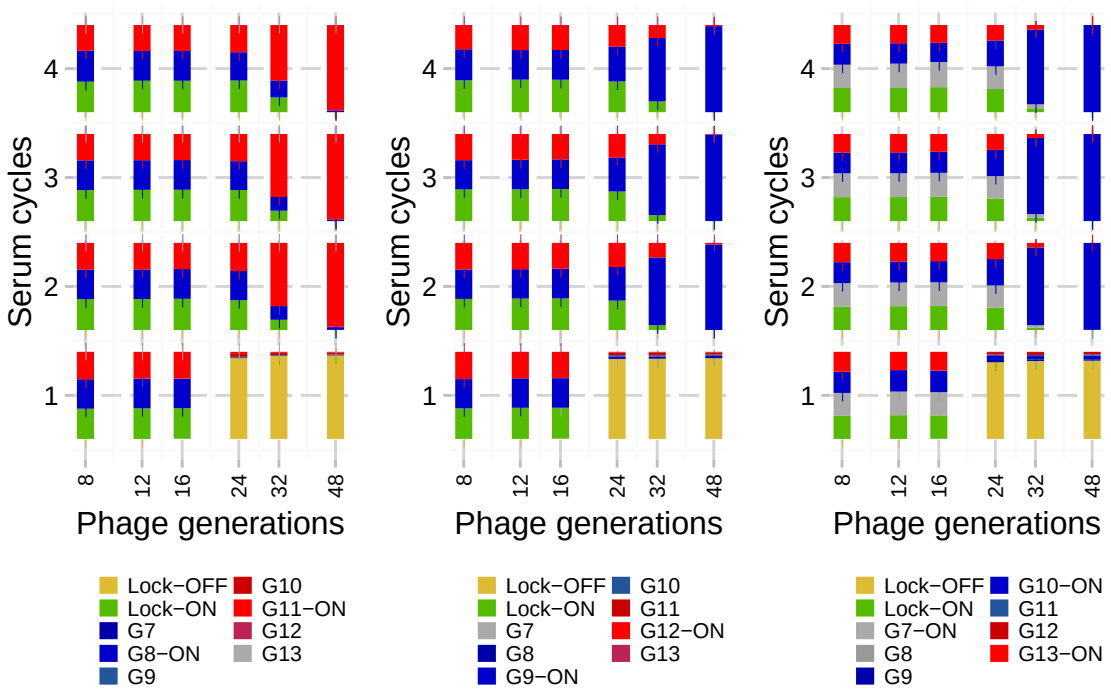

Figure 6.18: Simulation of *in vitro* cyclical selection experiments
This figure shows the predicted outcome of five cycles of selection under an empirically plausible range of periods of phage selection and repetitions of the serum selection cycle for the three possible ON length constructs G8-ON, G9-ON, and G10-ON shown from left to right.

suggest that it will not be possible to produce sets of selective conditions that favour the different ON lengths but incubating with phage for 24-48 generations (1-4 days) and a single cycle of serum selection will select for a locked-OFF variant over PV variants. Other combinations of conditions are expected to produce output populations similar to the input populations.

6.6 Conclusions

6.6.1 The size of non-selection bottlenecks qualitatively impacts outcomes

This data indicates that small bottlenecks will produce stochastic differences in major phasotypes between output populations, whilst intermediate bottlenecks

disrupt population structure, and the largest bottlenecks allow diversity to accumulate. These effects occur across a range of mutation rates and number of genes simulated. Comparison to *in vitro* experiments reveals close correspondence of simulated and experimental models with smaller bottlenecks but poorer convergence for large bottleneck sizes.

6.6.2 Changes in selective conditions will favour different phase variable or non-phase variable tracts

Despite differences in the mutational structure of *C. jejuni* and *H. influenzae*, this data shows that the range of conditions that favour phase variation is similar between *in silico* models of the two bacteria, however whereas the mutational patterns in *H. influenzae* drive the evolution of progressively shorter tracts, in *C. jejuni* the directional differences in mutational rates drive tract lengths towards intermediate values. The favoured length in most cases depends on the period of stability rather than the strength of selection once selection is sufficient to favour PV over non-PV. This model predicts that experimentally plausible conditions will favour PV over non-PV variants but that the same PV variant will be favoured in all conditions.

Chapter 7

Discussion

7.1 Summary of main findings

This work began with a survey of the prevalence of SSR mediated phase variation in *Campylobacter*s. This survey demonstrated that there are large numbers of SSR mediated phase variable genes across a broad range of *Campylobacter* species. The majority of these genes are involved in surface modification which accords with existing evidence on the functional roles of phase variable genes. Restriction/modification systems also featured prominently. Grouping these genes by homology indicated that the majority of PV genes identified are part of the accessory genome. Only a small number of homology groups are widespread among the species, however there is also evidence of a core phasome which is shared between most representatives of each species. The dataset used is not ideally suited to assess this core phasome because of inconsistent representation of the different species but it was still possible to identify a tentative core phasome for *C. jejuni*, *C. coli*, *C. fetus*, and *C. lari*. There was some overlap between the core phasomes of *C.*

jejuni, *C. coli*, and *C. fetus*, with the flagellar-modifying *cj1295* group found in all three species. There were also big differences in the size of these core phasomes: *C. lari* has just two core phasome homology groups, whereas the *C. fetus* core phasome contains 27 homology groups. The proportion of the total number of homology groups composed of the core phasome is also very different between species and this may reflect differences in the host or environmental range of these species. Despite the relative abundance of restriction/modification systems among the homology groups identified, none were found in the core phasomes of these four species.

Application of Phasome*It* to a larger dataset of host associated – but incomplete – genomes showed that there is an association between clonal complexes and the phasome, but that any host associating signal is swamped by this association with clonal complex. Sub-dividing by clonal complex allows a host association signal between the phasome and cattle-associated isolates of *C. coli* ST-828 complex to be identified which suggests that particular phase variable genes may adapt some isolates to particular hosts.

The major theme of the rest of the work was the development of a cyclical selection assay allowing a population of *C. jejuni* to be subjected to alternating selective bottlenecks forcing the population into opposite expression states of a particular phase variable gene. The target phase variable gene was *cj1421c* and it was demonstrated that phage F336 will select for the OFF state of this gene in a Δ*cj1422c* strain, whilst human serum will select for the ON state. Both of these selective agents produced output populations of $\geq 95\%$ of the selectively favoured state from input populations of $\geq 95\%$ of the opposite state. Furthermore, it was demonstrated that phage F336 and pooled human serum could be combined to produce a complete cycle of ON $\rightarrow$ OFF $\rightarrow$ ON in *cj1421c* using the Δ*cj1422c* strain. However, the efficacy of the serum selection depends on the source of that serum and a second sample of pooled human serum proved to have weaker selective effect and failed to clear

the bacteriophage from the population inhibiting the selective step. A model of this form of cyclical selection was also developed and revealed that a broad range of selective conditions could favour the evolution of phase variation. Running this model in the absence of selection demonstrated that neutral drift driven by the differential mutational patterns of the tract lengths leads to a population heavily dominated by the G9 (48.8%), G10 (27.1%), and G8 (14.0%) tract lengths. This neutral pattern is important in determining the behaviour of the model even under selection. This model also indicated that the favoured phase variation rate, and thus tract length, depends on the strength of selection and length of exposure but that this is likely dependent on comparative rates of loss through phase variation from ON states to adjacent OFF lengths. Fitting the model to experimental data indicates that using the cyclical selection assay with experimentally plausible conditions is likely to favour a G9-ON state over a G12-ON state.

A second model, created to simulate non-selective bottlenecks, revealed that the impact of non-selective bottlenecks is highly dependent on bottleneck size. Small bottlenecks are able to disrupt the population structure and produce heterogeneity between outcomes, whereas large bottlenecks allow an increase in diversity and produce similar outcomes. Such small bottlenecks are likely to be common in the lifecycle of these bacteria during host invasion and transmission, and thus this provides a mechanism for the proliferation of distinct populations that may have different disease implications. For example, separate populations could be more or less likely to result in disease or could have raised probabilities of rare post-infection complications.

7.2 Phase variable genes are enormously diverse

The majority of phase variable genes discovered are found only in a small number of genomes. Although in some of these cases there are non-phase variable

homologues of these rare phase variable genes found in other genomes most have no homologues in other genomes. Thus these genes form part of the accessory genome of these species. This finding is surprising because it is thought that in order for phase variation to evolve there must be selection for both the ON and OFF states intermittently applied (Beaumont *et al.*, 2009, Libby and Rainey, 2011, Palmer *et al.*, 2013, and also see findings in section 6.4). If so many phase variable genes evolved separately in these lineages it suggests that individual strains must be subject to strong fluctuating selective conditions. However, an alternative perspective is that the environmental conditions that favour the OFF state also allow for high frequency loss of phase variable genes and thus the high proportion of phase variable genes in the accessory genome reflects frequent loss of these genes. The existence of non-PV homologues of many PV genes is the converse of this effect since it suggests that the environments selecting against expression of these genes are themselves rare or, alternatively, may be a measure of the degree to which the absence of these genes is tolerated and represent a non-advantageous development of phase variation. A possible candidate for this kind of non-advantageous evolution of phase variation is *cj0275* (*clpX*) which appears only in the ON state during chicken passage experiments (Lango-Scholey *et al.*, 2016).

One consequence of the phasome being predominately composed of genes in the accessory genome is that searching for homologues of known PV genes will be ineffective at identifying the full range of phase variable genes present in a genome and, instead, it is necessary to search for the variable regions themselves and then identify the associated gene.

7.3 The functional roles of PV genes

Out of 536 homology groups identified in the complete genome collection, 176 contained no members with information about their function but in the majority

of cases it was possible to assign putative function to the homology group. The characterisation of functional role is dependent on the annotation data applied to the genomes, and this is based – in most cases – on comparisons to genes of known function and the assumption that protein sequence similarity implies functional similarity. This same assumption underpins the collection of homologous genes into groups that are presumed to share functional similarities. The degree of validity of this assumption is unclear. Pessimistic estimates suggest that as few as 30% of functional deductions based on sequence similarity are correct (Rost, 2002), however this analysis applied no length cutoff to the matches and thus allowed high similarity matches between small subdomains of the proteins. Since Phasome*It* applies bi-directional coverage limits to the accepted matches, it is likely the accuracy of the matches is closer to that in more optimistic estimates that find good accuracy of functional assignment down to low levels of sequence similarity (40%) where these matches cover the length of the protein (Devos and Valencia, 2000, Wilson *et al.*, 2000, Whisstock and Lesk, 2003).

A related concern is that the original annotation data itself may be in error, and that these errors can propagate between automatically annotated genomes. These annotation errors have been estimated to influence between 8% and 37% of all annotation depending on how stringent a criteria is applied (Brenner, 1999, Devos and Valencia, 2001, Andorf *et al.*, 2007, Schnoes *et al.*, 2009). Taken together these results suggest a significant degree of caution is required in interpreting the homology groupings identified by Phasome*It* and their functional assignments.

7.3.1 Most PV genes are surface associated

The majority of these homology groups involve modification of surface structures, and the three largest homology groups are all surface modifying. This is consistent with previous results on the function of phase variation which suggests

that it is primarily involved in altering surface structures to prevent recognition by other molecules and organisms (van der Woude and Bäumler, 2004, Moxon *et al.*, 2006). This diversity in surface structure may mediate interactions with the immune system or bacteriophages, as well as having possible roles in motility or agglutination.

7.3.2 PV restriction/modification systems are common

A major class of homology groups identified was restriction modification systems. Phase variable R/M systems have been previously identified in a range of species (Ryan and Lo, 1999, De Bolle *et al.*, 2000, Manso *et al.*, 2014, Anjum *et al.*, 2016). The significance of phase variation in these systems is proposed to come from two sources: (1) it may prevent build-up of resistant phage (Bayliss *et al.*, 2006), and/or (2) there may be fitness benefits associated with the disabling of the system, e.g. by restricting plasmid transfer (Donahue *et al.*, 2000, Seib *et al.*, 2002), and/or (3) the methylation patterns may control expression of a range of genes. Because the selectivity of the system is controlled by a small region of the protein, the target recognition of even systems with high homology may be different and thus the individual members of these R/M homology groups may actually be quite different to each other in terms of their target sites as is the case with *cj0031* and its homologues (Anjum *et al.*, 2016, Murray *et al.*, 2012). If methylation is involved in further gene regulation even small differences in the recognition site could have dramatic effects on the regulated genes.

Interestingly, despite their wide occurrence, no R/M systems were found in the core phasome of the species analysed. This may reflect the need for diversity in these systems in order to benefit from their restrictive effect against phage colonisation. Bacteriophages, like other viruses, have extremely high mutation rates and thus can rapidly mutate recognition sites to avoid restriction modification systems and

this may impose strong selective pressure for diversity in R/M systems. However, the placement of R/M systems into different homology groups is unlikely to result from differences in target recognition domains. This is due to the low homology cut off used in forming homology groups, and so it is not clear a need for simple diversity in recognition sites would drive this lack of conservation in R/M system homology groups. Additionally, the *cj0031* and *cj1051c* groups have many non-PV homologues and so represent localised selection for phase variation rather than for systems themselves but the other identified homology groups are sparsely represented in both PV and non-PV form.

7.4 Comparisons between the site of SSRs in PV genes and their non-PV homologues

How phase variable loci initially evolve remains a currently open question. Four possibilities are: (1) equivalent amino acid coding sequences mutate to produce an SSR by neutral drift (i.e. three consecutive glycines become 9Gs); (2) A shorter tract lengthens to become a variable tract; (3) recombination of the region around the SSR transfers the tract to a new gene; or (4) random mutation happens to generate a mutable sequence. This is particularly relevant to *Campylobacter* because the low G/C content of these genomes means that random chance is unlikely to produce tracts long enough to become variable. Phasome*It* was not designed to answer this question however comparisons between PV and non-PV loci (q.v. figure 3.11) are potentially relevant. This provides evidence of disrupted tracts (i.e. GGGGGGAGGG) where a point mutation could potentially create a phase variable tract, or conversely this sequence could have resulted from a previous PV tract; and this second possibility seems more parsimonious. There are also shorter tracts (e.g. AAAAGGGGG) but it is not clear what process might convert these to longer tracts as mutation rates are low. However, it is possible that if positive selection is absent long enough the

sequence could undergo successive mutations to lengthen the repeat sequence. Finally, there are genes where the sequence is entirely dissimilar which could potentially undergo (3) or (4) to produce a new variable locus. Recombination would be expected to produce multiple tracts with similar regions surrounding the SSR however visual inspection of the dataset does not reveal any evidence of this with exception of pairings such as *cj1421c* and *cj1422c* which share a large (around 1000bp) region of homology. This region appears to undergo frequent recombination as the regions in pairs of genes in different strains are much more similar to each other than they are to the regions in other genes regardless of the similarity of the rest of the gene. These regions have also been observed to spontaneously recombine under laboratory conditions (McNally *et al.*, 2007).

7.5 Most *Campylobacter* species have a "core phasome" of phase variable genes common to members of that species

One of the novel findings of this analysis of *Campylobacter* is the identification of a core phasome associated with distinct *Campylobacter* species. The homology groups in the core phasome are not necessarily present in PV form in all members of that species but are there more often than not. The genes in these homology groups, therefore, are likely to be broadly beneficial to these species in a PV form and thus offer an insight into the selective pressures acting on these species. Four species in the dataset were present in sufficient number for a putative core phasome to be identified: *C. jejuni*, *C. coli*, *C. fetus*, and *C. lari* and there are dramatic differences within this set of core phasomes. The most obvious is in the size of the core phasomes: 2 for *C. lari*, 10 for *C. jejuni*, 12 for *C. coli*, and 27 for *C. fetus*. Although the proportion of homology groups that is conserved are actually similar for *C. coli*

and *C. fetus* as *C. fetus* genomes typically contain around twice as many homology groups as those of *C. coli*. There are fewer homology groups in *C. lari* overall but the 2 in the core phasome represent a smaller proportion of these than any other species.

Of more interest are the genes that compose these homology groups. Despite the similarities between *C. jejuni* and *C. coli*, there are only 4 homology groups shared between them: *cj0170*, *cj1295*, *maf1*, and *maf7*. Intriguingly what is known about the homologues present in *C. jejuni* strains NCTC 11168 and 81-176 suggests all of these homology groups are involved in motility and operate by modifying the flagellum (Karlyshev *et al.*, 2002, van Alphen *et al.*, 2008, Hitchen *et al.*, 2010, Artymovich *et al.*, 2013). Of these shared groups, the *cj1295* group is also shared with *C. fetus*. Many of the genes present in the large core phasome of *C. fetus* are unfortunately lacking in useful annotation data, however SAM-dependent methyltransferases feature highly. SAM is an ancient molecule, probably dating back to the common ancestor of all cellular life, and the SAM-dependent methyltransferases are a large and diverse group of transferases found in all branches of the tree of life (Kozbial and Mushegian, 2005).

7.6 Tract length frequencies appear to follow neutral patterns, but the proportion ON does not

The model predicts that, using the central estimates of mutation rate, the 9 length tract will dominate the population in the absence of selection, with 48.8% of the population (see section 6.4.2). This distribution results from differential patterns of insertion and deletion between tract lengths as well as the frequency of mutation. The observed tract length frequencies are close to this neutral prediction suggesting that tract lengths are driven primarily by neutral forces. Alternatively it could be

that 9G is the most common tract length as it most easily evolves from an amino acid sequence of three glycine residues. However, since the proportions of putative ON lengths are also very close to the predicted neutral drift proportions this may reflect the optimal proportion ON in the entire population. This should result in the proportion of putatively ON genes following the predicted proportion of ON by tract length but this is not the case and the percentage of genes in the ON state is higher than neutral drift would predict. This higher proportion suggests that the strains have undergone selection for expression of the ON states or, alternatively, that selection for the OFF state is typically weaker than selection for the ON state. Curiously, this proportion ON is lower in the host attributed set of incomplete genomes than in the set of complete genomes (although still higher than predicted by neutral drift). There is no obvious reason why this should be the case and it may be that differences in assembly pathways that the two genomes have undergone produce artefacts in the tract length data.

7.7 Findings from host-associated dataset

7.7.1 The phasome is associated with ST complex

The presence or absence of particular PV homology groups follows patterns of ST-Complex. However, even within these groups there is a large variety of PV homology groups. Most likely this within complex variation represents horizontal transfer of these phase variable loci. Although not investigated in this analysis the core phasome concept could be extended to the grouping by ST-Complex and the nature of these ST-Complex core phasomes may be relevant to the life-cycles of these groups.

7.7.2 Limited evidence for the association of homology groups with particular host attribution

There is some evidence of a link between the phasome and the host colonized by the organism but this signal is much weaker than the association with ST-complex and only strongly supported in the ST-828 complex of *C. coli*. This may reflect a lack of any role of the phasome in host specialisation, alternatively it may represent either a diversity of available adaptations or be a consequence of the bacteria moving between hosts. The host attribution is nothing more than the host from which the bacteria was finally isolated, and may not represent the life history of the bacteria. As *C. jejuni* and *C. coli* are known to move between hosts; it may be that an isolate obtained from chicken is actually adapted to cattle, and vice-versa. This is most clearly seen in the "dead-end" human disease isolates which cluster with isolates from all other hosts.

Alternatively the large numbers of rare PV groups may play a role in host adaptation. If these groups are able to substitute for one another in adapting the bacterium to a particular host this would dilute any association signal and thus not be apparent in the dataset. Selection on phase variable loci may involve either specific or non-specific interactions and genes *cj1421c* and *cj1422c* provide an example of this. The interaction with phage is specific; i.e. phage attachment is dependent on the precise positioning of the MeO*P*N group as illustrated by the fact that phage attachment and invasion will proceed in a *cj1421c*-ON/*cj1422c*-OFF variant but not in a *cj1421c*-OFF/*cj1422c*-ON variant. This is despite the attachment of the same moiety to closely separated parts of the same molecule. In contrast, serum selection is non-specific because either MeO*P*N transferase switching to the ON state will increase resistance to pooled human serum. If the required adaptation is non-specific then it may be adequately fulfilled by genes from a number of homology groups. Alternatively, micro-environmental factors in host adaptation (e.g. phage

colonisation with different strains, or differences in the microbiome) may strongly influence the phasome and swamp broader host signals.

7.8 Phasome*It* is applicable to a wide range of bacteria

In this thesis, I have concentrated on the phase variable genes present in *Campylobacter* but Phasome*It* is able to detect and compare phase variable genes in any species that uses the same mechanism of variation. This covers a wide range of bacteria with a diverse range of biological niches (Lin and Kussell, 2012).

Because Phasome*It* only captures one mechanism of generating phase variation, this survey is necessarily incomplete and may not capture the whole phasome of the organisms studied. It is known that in *C. fetus*, there is at least one other mechanism of phase variation. This organism exhibits stochastic re-arrangements of the coding sequence for a surface layer protein that covers the organism in a monomolecular array, produced by recA-dependent re-assortment of a series of nested repeats to produce a wide range of possible sequences (Dworkin and Blaser, 1997). It is unclear how widespread these alternate mechanisms of phase variation are in the *Campylobacter*s but, to date, no such mechanisms have been identified in the most closely studied species (*C. jejuni*) whereas potentially phase variable repeat tracts were identified in every genome looked at which suggests it is the more common mechanism.

An additional difficulty is determining what impact the PV loci identified have. Where the tract is located within the reading frame and close to the start, this gives considerable confidence that it operates via a frameshift mechanism as described in the introduction. However, 24.6% of the identified SSRs are located outside of any

open reading frame. In other species the positioning of the tract between promoter elements means that differences in tract length can impact the efficiency of binding of transcription factors and thus produce differential expression levels (van der Woude and Bäumler, 2004, Moxon *et al.*, 2006, Alamro *et al.*, 2014) and these SSRs may be influencing gene expression in the same way.

Another difficulty arises if the tract is located closed to the end of the protein. In *C. jejuni* strain NCTC 11168, the gene *cj0170* has the tract located close to the end resulting in up to 22 amino acids difference in product length, however it is not the longest product that is functional but the shortest (Artymovich *et al.*, 2013); the exact mechanism for this is currently unclear. A second gene, *cj0045c*, also has the tract located close to the end, but in this case the longest product overlaps the start codon of the adjacent gene *cj0044c* and appears to affect transcription levels of this adjacent gene when it does so (Bayliss group, unpublished data).

Taken together these difficulties mean that the automated approach is imperfect in analysing the phasome of these species, as it may both contain spurious non-PV genes and miss genes that are phase variable by other methods, and misattribute the ON state of found genes. These cases therefore suggest caution is necessary in interpreting the results. However these limitations do not undermine the major findings on phase variation in *Campylobacter* nor limit the application of Phasome*It* to other species but rather underline the need for bioinformatic analyses to work hand-in-hand with experimental work.

7.9 Both the ON and OFF states of gene *c1421c* can be selected for *in vitro*

7.9.1 Phage is effective as a selective agent *in vitro*

In every experiment incubation in the presence of phage F336 produced a selective sweep for the ON state of *cj1421c* thus demonstrating its effectiveness as a selective agent for the cyclical selection assay. There was no evidence of selection for any gene other than the two MeO*P*N transferases except in a single replication where a mild change in *cj1426c* occurred. This suggests that the reduced EOP observed for *cj1426c* is insufficient to produce strong selective pressure for the ON state of this gene and that it is unlikely, therefore, to interfere with the cyclical selection assay.

7.9.2 MeO*P*N is protective against human serum

Serum samples from different volunteers showed different levels of killing, suggesting that some component unique to individuals is responsible for the killing effect. The most likely candidate is a *Campylobacter* specific antibody since there is evidence that the killing effect is mediated by the classical complement pathway (van Alphen *et al.*, 2014) in *C. jejuni* strain 81-176. It is possible that the addition of the MeO*P*N group is protective because it blocks the particular epitope recognised by this antibody. This interpretation seems unlikely because the effect was uniform across samples. It is unlikely that this particular region of the CPS would be uniformly, uniquely, and consistently specified as an antibody target. A more parsimonious explanation, therefore, is that the MeO*P*N groups block the action of the killing itself rather than the antibody. A second line of reasoning that leads

to the same conclusion is that the serotype of NCTC 11168 is not known to vary with the state of the two MeO*P*N transferases. MeO*P*N is a highly negatively charged moiety and this may play a functional role in the interaction with serum but whether the group binds to an acute phase serum protein, prevents antibody deposition, or acts through another mechanism is unknown (van Alphen *et al.*, 2014). Phosphoramidate groups without the O-methyl group of the MeO*P*N moiety of *C. jejuni* have been identified in *Shewanella spp.* (Vinogradov *et al.*, 2008) and *Xanthomonas campestris* (Silipo *et al.*, 2005). In *X. campestris* this similarly highly negatively charged group has been shown to be protective against the innate immune system of *Arabidopsis thaliana* and here evidence was found that the charge influences binding to plant receptors (Silipo *et al.*, 2005).

7.9.3 The protective effect of MeO*P*N expression is sufficient to select for the ON state

The concentration of pooled human serum required to produce a complete selective sweep for the ON state of the MeO*P*N transferases was variable between collections of serum and between individual serum samples, however it was possible to produce a total selective sweep for the ON state with both pooled human serum samples. There was also no evidence of any selective effect on the other PV loci of NCTC 11168.

7.10 Viability of a selective cycle

The work in chapter 5 demonstrates that a cyclical selective assay can be constructed using a combination of the selective effects of phage F336 and pooled human serum and there is no indication that this assay will influence any off-target gene

(i.e. any gene other than *cj1421c*). This thus forms an effective base for future experimentation.

The phage selection step of this cycle has proven to be reliable in a range of condition and with a range of starting populations and whilst different concentrations of serum were required, the serum selection step worked with pooled serum collected from different sources. These differences in selectivity between the two pooled serums will require the serum selection step to be re-standardised between collections before use. Another complication in the deployment of the cyclical selection assay is the failure of the second pooled human serum to clear the bacterial suspension of bacteriophage. Without successful removal of bacteriophage the ON selection step cannot be verified. Accordingly pooled serum to be used in the sample needs to be tested for bacteriophage clearing effect before use in the cyclical assay or an alternative used. One possibility is to raise anti-serum specific to bacteriophage F336 and use this to remove the phage before proceeding to the pooled human serum selection step.

7.11 Modelling of selection/non bottlenecks

Pathogenic bacteria experience tight bottlenecks during spread within hosts, and transmission between hosts (Gerlini *et al.*, 2014, Meynell and Maw, 1968, Rubin, 1987). These bottlenecks can impose a significant reduction in population diversity and one fitness benefit of phase variation may be re-establishment of diversity in these populations after the imposition of a population bottleneck.

7.11.1 Single cell bottlenecks as potentiators of disease heterogeneity

Single cell bottlenecks produce a distinctive bimodal distribution in population divergence of phasotypes that is absent in larger bottlenecks. Intermediate bottlenecks show decreasing levels of finer structure as the distributions converge towards the patterns of the larger bottlenecks. This bimodal pattern likely results from rare phasotypes being included in the bottlenecked population which then come to dominate the new population. This results in a dramatically different population structure in a close parallel to the founder effect in classical evolutionary models (Barton and Charlesworth, 1984). Further examination of the divergent populations in single cell bottlenecks indicated that there was divergence from the initial population but also from each other. A total of 60 divergent populations were found in 100 simulated output populations in the *in silico* model, which was mirrored by 8 in the 20 experimental populations. Each of these diverged populations is dominated by a phasotype which differs from the major phasotype of the inoculum in the expression state of one or more genes.

7.11.2 The high levels of divergence caused by single-cell bottlenecks has the potential to impart stochastic effects on the outcome of infections

If specific phasotypes are responsible for disease progression, then only populations where these phasotypes are produced as the dominant type may cause disease (figure 7.1). The net result would be a stochastic appearance of disease in a population even though the pathogenic organism is widely prevalent. High carriage to disease ratios are features of meningitis due to *Haemophilus influenzae*

and *Neisseria meningitidis* (Coen *et al.*, 1998, García-Rodríguez and Fresnadillo Martínez, 2002, Caugant *et al.*, 2007) two pathogens that contain multiple phase-variable genes with known or putative roles in disease-related phenotypes such as immune evasion (Moxon *et al.*, 2006). Similarly, there are frequent infections with *C. jejuni* strains capable of expressing epitopes that may induce the auto-antibodies responsible for neuropathies and yet these post-infection sequelae are rare. These pathogens are subject to small bottlenecks due to passage between compartments (e.g. nasopharynx to bloodstream) or during transmission (e.g. infected chicken through the food chain to contaminated food product). Divergence in phasotypes (i.e. phase-variable gene expression patterns) due to small bottlenecks may be a component of the low disease frequencies of such pathogenic bacteria.

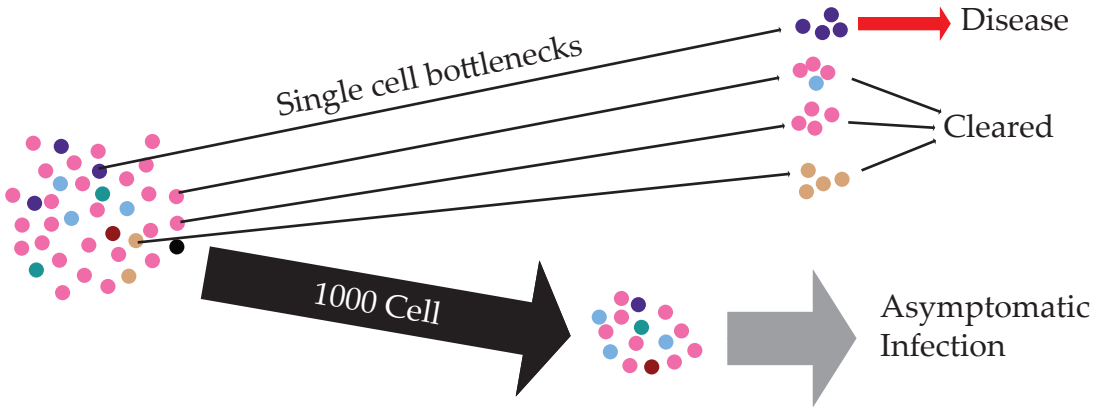

Figure 7.1: Impact of stochastic changes in population phasotypes on disease during host-to-host spread
This diagram depicts the potential effects of non-selective bottlenecks altering the population structure of a bacterial species. The initial population in a transmitter host contains a variety of phasotypes (represented by different colours) but a single major phasotype (pink). Non-selective bottlenecks may disrupt this initial population structure, with single-cell bottlenecks potentially capturing a minor phasotype and producing a new population with this as a major phasotype. One of these phasotypes has the potential to cause disease in a new host whilst the others are cleared resulting in a stochastic effect on the occurrence of disease. In this example, only 1 in 4 hosts exhibits disease. In the case of a larger 1,000 cell bottleneck, the diversity of the initial population is retained enabling the population to rapidly establish an asymptomatic infection. Figure based on Christopher D. Bayliss (personal communication).

7.11.3 Maintenance of population diversity through relatively small bottlenecks

While very small bottlenecks have a dramatic reductive effect on phasotype population diversity, larger bottlenecks preserve much of the phasotype diversity within a population. This effect begins to dominate even at relatively small bottleneck sizes numbering in the 100s of cells and is also observed even with highly diverse starting populations. Because each bottleneck preserves the diversity generated during the growth phase of the population, there is a correlation between divergence and diversity as the population diverges from the original by becoming more diverse. This preservation of initial diversity highlights another alternate pathway whereby phase variation could contribute to disease or host adaptation by a bacterial pathogen. In situations where the transmitted population is in the range of 100-1,000 cells, the diversity generated in one host individual or compartment would be maintained through the transmission event. This preservation of diversity may be important as it may facilitate adaptation in the new environment. Thus, for example, the phase-variable genes of a *C. jejuni* strain may have reached the optimum fitness state in the caeca of one chicken, preservation of these states through a transmission bottleneck will then facilitate rapid colonization of the next chicken and hence lead to rapid spread through a flock. Preservation of diversity may also preserve the population in a state ready for random insult, and avoid local extinction events associated with less diverse populations.

7.11.4 Predictions of the *in silico* model match experimental results for smaller bottlenecks

The model was not fitted to the experimental data but rather parametrised using known values from earlier experiments to avoid circular pathways of reasoning

and to ensure relevance to other systems. Despite certain simplifying assumptions, the major features of the experimental data for the bottlenecks from size 1 to 100 were reproduced in the *in silico* model, suggesting that this model captured the underlying behaviour of the biological system. At the 1000 cell bottleneck, the correlation between experimental and *in silico* data was weaker than with small bottlenecks, mainly resulting from a greater than predicted divergence. Selection does not seem to be responsible for this effect as there were no specific genic or phasotype differences. An alternative cause are the assumptions made in creating the model. Several of these assumptions are known to be simplified from the biological realities. These include all genes switching at the same rates in both directions; a single OFF state; and identical PV rates for all loci. Minor variations in the mutation rates and directions of switching due to these assumptions may generate minor alterations in population structure that accumulate in large bottlenecks but are disrupted by the stronger effects of small bottlenecks. Improved experimental evaluation of switching rates may be required to further dissect the interplay between mutability of PV genes, bottlenecks and population structure.

7.12 Modelling of repeated selective bottlenecks

7.12.1 Cutoff for transition between high and low rate changes is usually independent of the strength of selection

In the case of the G11-G8 transition and G10-G7 transition the point of transition occurs at a fixed epoch length T. This indicates that rather than being driven by the selective force itself, this transition is the result of neutral drift that occurs as a result of differing rates of loss of the ON state to the neighbouring OFF states because of the relative mutation rates of the two tract lengths. In contrast, the

G12-G9 transition not only occurs in an s-dependent manner but also shows a more complex pattern of transition from G9-G12-G9. This pattern is likely to result from an interaction of neutral and selective forces. At low T and high s, the G9 tract is favoured by neutral forces but as the time increases the larger switched populations present under the higher switching rates of the longer tracts (G11-G13) compared to the smaller switched populations of the shorter tracts (G8-G10) allow the longer tract to dominate the population. As the time increases, the shorter tract again begins to dominate as the length of T allows these shorter tracts to out compete the less stable tracts.

7.12.2 Experimental relevance of *in silico* modelling of the cyclical assay

As well as this broader look at the interaction of alternating selection and tract length/phase variation rate, I also used the model to predict the outcome of experiments using the cyclical selection assay described in chapter 5. These predictions indicate that it will be possible to carry out experiments over reasonable timescales that either favour a locked-OFF mutant (or, equivalently, a simple double knockout Δ*cj1421c*/Δ*cj1422c* strain) or a PV mutant depending on the conditions chosen. Unlike the experiments of Acar *et al.* (2008) carried out with *Saccharomyces cerevisiae* it does not appear to be possible to tune experimental conditions to select between faster and slower switching rates. This likely reflects differences in both the switching rate of *C. jejuni* (which is about an order of magnitude lower than the yeast constructs used) and the selective agents used.

7.13 Progress towards aims

The work presented here represents substantial progress towards the goals set out in section 1.8 but more work needs to be done. Phasome*It* and the analysis of *Campylobacter* phasomes presented in chapter 3 have substantially increased our knowledge of the conservation, and diversity, of phase variable genes in *Campylobacter*. The packaging of this analysis into a convenient and widely applicable program allows easy extension of this analysis to other species. However, the significance of the revealed diversity in potentially PV genes has yet to be fully interrogated and there is great potential for further analysis of the generated datasets. The analysis of host-associated isolates is suggestive of links between the phasome and particular hosts, however larger datasets and experimental work are needed to establish the significance of these links.

The aim of creating a cyclical assay and using it to interrogate questions regarding the biological significance of phase variation rate is partially complete. This work has demonstrated the viability of phage F336 and human serum as selective agents in Δ*cj1422c* strains of NCTC 11168 and established the specificity of the selective effect on *cj1421c*. However, inconsistent selective effects from the serum and the lack of a reliable phage removal step prevented the full deployment of the cyclical assay to its experimental conclusions.

Together with the experimental cyclical assay, this work also includes a model aimed at the same question. Results from this model give significant insight into the range of conditions likely to favour different phase variation rates. These data can be compared to that of Palmer *et al.* (2013), and despite the dissimilar mutational mechanics of *C. jejuni* and *H. influenzae* give comparable results for the conditions that favour phase variation. Because the cyclical assay was not progressed to its

final stages, it was not possible to compare these predictions to real experimental results.

7.14 Future work

The data generated by Phasome*It* is an enormously rich and detailed dataset and the findings reported in chapter 3 only scratch the surface of the information in this dataset. Particularly important will be linking the findings reported, and the variation observed, to the life cycles of these organisms. As with any bioinformatics project, there is a need to relate hypothesises resulting from the analysis to experimental investigation.

Phasome*It* is also suitable for application to any sequenced bacteria that uses SSR-mediated phase variation. Other members of the Bayliss group have already applied it to *Neisseria meningitidis* but this mechanism is present in many more bacteria. The degree of knowledge of the phasome of these bacteria is highly variable and so the application of Phasome*It* to these bacteria has the potential to significantly increase our understanding of phase variation in general.

In this study, I have used serum selection as an experimental technique rather than seeking to understand the underlying mechanisms. This work, building on the study by van Alphen *et al*. (2014), supports the protective mechanism of MeO*P*N against serum and the attribution of the action of killing to particular elements of the classical complement pathway, however it has not sought to provide proof of this pathway. The isolation of particular antibodies, confirmation of binding, and investigation of the mechanism of action are all interesting avenues for further investigation of this link which fell beyond the scope of this project.

The development of a cyclical assay lays the groundwork for the use of *C. jejuni* as a model organism to study aspects of phase variation and, in particular, the fitness and significance of phase variation rate under different selective conditions. The first step will be to identify a reliable method for phage removal or source new pooled serum that is able to remove the phage. Once this hurdle is overcome the theoretical results from the model can be tested. Particular experiments to perform include competition experiments with phase variable and non-phase variable mutants, and the competition of G9 and G12 ON strains. The construction of mutants with different ON states and competition between these is also potentially possible and would open up the available range of different mutation rates to compare. Because selective bottlenecks imposed on one gene impose a non-selective bottleneck on the other genes, the state of other PV loci in these experiments could be related to the predictions of the non-selective bottleneck model.

Appendix A

Primers used for 28-locus-CJ11168 PV-analysis

Mix	Gene	Primer name	Sequence (5′→3′)
A	*cj0031*	cj0031-fwd-FAM	GGCTTTGATCTCATCATCGG
		cj0032-rev	GCAAAGCTTCCCCATATCCT
	cj0045c	cj0045-fwd-FAM	TTTTACACTAGAACACAGAAG
		cj0045-rev	CCTTAAAGTGCGAAAAATGTG
	cj0628	CapA-fwd-FAM	TATTTCTAATGATGGGCAAC
		CapA-rev	GAACGAACATTTACACCCAT
	cj0685c	cj0685-fwd-FAM	GATAGCGAATATAACCTCTAAATTC
		cj0685-rev	GAAGAAATCCGCCAATCAAAG
	cj1325	cj1326-fwd-FAM	CTTTTGGAATAGATATAGTTCC
		cj1326-rev	TTAGAGGTATGTAGTAAAGAC
	cj1342	cj1342-fwd-FAM	TTGGCAATCGTCCTCAAACC
		cj1342-rev	GCCAAATGCGCTAAATATCC
B	*cj1318*	cj1318-fwd-FAM	TCCGTGCGTCCTTCTTTTGGAC
		cj1318-35-rev	GTTTGCAACTCTTTAATGGG
	cj1420c	cj1420-fwd-FAM	GCTAGTTCTTTCCATTGGAC
		cj1420-rev	CTACAATGTGGCGAGGATTC
	cj1426c	cj1426-fwd-FAM	TATAGCCGATCCACAAGG
		cj1426-rev	GATATAACTTTGCCCGCCAC
	cj1429c	cj1429-fwd-FAM	ATGGAGATGGTGGTTATGTG
		cj1429-rev	ACTATCCGAAACACCCAAAG
	cj1437c	cj1437-fwd-VIC	GTGCTAGGATGGAATTTGTG
		cj1437-rev	CAAACAAGGTGAAAACCTCC
C1	*cj0275*	cj0275-fwd-NED	ATTACTCGTGATGTAAGTGG
		cj0275-rev	AAACCTACAACTTTATCTCC
	cj1296	cj1296-fwd-NED	ATAAAGTGCATTCTAAAGGC
		cj1296-rev	CAGCAAAGGAAAAAATAGGG
	cj1306c	cj1306-fwd-NED	TTTATTCCTTCGCGTGGAGA
		cj1306-rev	AAAAATGATCGCCCTGCAT
C2	*cj0170*	cj0171-fwd-NED	TGGTTGTGGAAATGGAGTGC
		cj0171-rev	GCTCCTTCATTGCATAGTTC
	cj1144c	cj1144-fwd-NED	GATGTTGTGATTCTTG
		cj1144-rev	GTAGCAGCGTTTAGTG
	cj1335	cj1335-fwd-NED	CACAATTGGTTTATCCAAGG
		cj1318-35-rev	GTTTGCAACTCTTTAATGGG

Mix	Gene	Primer name	Sequence (5′→3′)
D	*cj0565*	cj0565-fwd	AATTTCACTTCCCCCTTGACT
		cj0565-rev-VIC	TTTTGCAACATCGCGTAGAA
	cj0617	cj0617-fwd-VIC	TGGTATAATGCAAGCTATGG
		cj0617-rev	AAATCAATACTCCAAGGAGC
	cj1139c	cj1139-fwd-VIC	GCAACTTCACCTTATATC
		cj1139-rev	TAAATTCTTTGTTGTTGTATTTTCC
	cj1305c	cj1305-fwd-VIC	CAACTTTTATCCCACCTAATGGAG
		cj1305-rev	AAAGCCGAACCCGAATTATC
	cj1321	cj1321-fwd-VIC	AAAAAGGAATGATGCGTTGC
		cj1321-rev	CCCGCTCCTATGATGATGAC
	cj1421c	cj1421-22-fwd	TTGGGTATTTAAGTTGGGGAAA
		cj1421-rev-VIC	TCAAAACCCATCTTTATCATTTTCT
E	*cj0046*	cj0046-fwd-NED	TCAAATACTGCAAGAGCAGG
		cj0046-rev	TAGAAGCATTAGGCGTGG
	cj0676	cj0676-fwd-NED	ATGCTTATTCCTAGTGCCTG
		cj0676-rev	TGCATTTAAACCCAAAGAATCC
	cj1295	cj1295-fwd-NED	TTCCTATCCCTAGGAGTATC
		cj1295-rev	ATAGGCTTCTTTAACATTCC
	cj1310c	cj1310-fwd-FAM	GAACAAATTATTCCTCTTATAG
		cj1310-rev	TCGAAATAAAATTCCCCTTGA
	cj1422c	cj1421-22-fwd	TTGGGTATTTAAGTTGGGGAAA
		cj1422-rev-NED	AATGATTTTGCTTTGCAGGAA

Primers from Lango-Scholey *et al*. (2016)

Six multiplexed PCRs are carried out with primers from the six mixes: A, B, C1, C2, D and E. Primers ending in 'FAM', 'NED', or 'VIC' have the named fluorescent marker attached, other primers are not marked. Note that *cj1318* and *cj1335* share a forward primer, and *cj1421c* and *cj1422c* share a reverse primer so there are only 54 primers in total, not 56.

Appendix B

Complete list of homology groups identified by Phasome*It*

Complete list of homology groups identified in the complete *Campylobacter* genome set, with automatically identified group names and functions, as well as the species they are found – in PV form – in one or more isolates of. Species names are abbreviated as follows:

Species	Abbreviation
C. coli	Ccl
C. concisus	Ccn
C. curvus	Ccu
C. fetus	Cf
C. gracilus	Cg
C. hominis	Cho
C. hyointestinalis	Chy
C. iquanorium	Cig
C. insulaenigrae	Cin
C. jejuni	Cj
C. lari	Cl
C. peloridis	Cp
C. sp.	Csp
C. subantarcticus	Csu
C. ureolyticus	Cu
C. volucris	Cv

Name	In species	Putative Function	# in group[1]
paeR7IM	Ccl Cj Cp	type II restriction endonuclease	12/12/56
cj0045c	Ccl Cj	Hemerythrin-like iron-binding protein	43/43/45
cj0046	Cj	pseudogene (putative sodium:sulfate transmembrane transport protein)	1/1/1

Name	In species	Putative Function	# in group[1]
cj0067	Cj	chlorohydrolase	18/18/73
cj0170	Ccl Cin Cj Cl Csu Cv	SAM-dependent methyltransferase	47/52/54
clpX	Cj	ATP-dependent Clp protease ATP-binding subunit ClpX	13/13/77
a911_t08342	Cho Cj	No annotation data	0/21/75
cj0565	Cj	pseudogene (conserved hypothetical protein)	0/5/7
maf7	Ccl Chy Cin Cj Cl Cp Csp Csu Cv	carbonic anyhydrase	200/203/233
cj0628	Cj	putative lipoprotein	19/23/35
kdpA	Cj	pseudogene (potassium-transporting ATPase A chain)	10/10/32
cipA	Ccl Cin Cj Cl Csu	Invasion protein CipA	36/36/39
hxuB_1	Cj	Heme/hemopexin transporter protein HuxB precursor	0/25/36
wlaN	Cj	beta-1,3 galactosyltransferase	11/11/15
cj1144c	Cj Cp	hypothetical protein Cj1144c	16/16/18
cj1295	Ccl Cf Chy Cin Cj Cl Csp Csu Cv	hypothetical protein (DUF2172 domain), putative M28 family zinc peptidase	64/65/72
cj1296	Ccl Cin Cj Cl Cp Csu Cv	aminoglycoside N3′-acetyltransferase	29/30/92
maf1	Ccl Cin Cj Cl Csu Cv	motility accessory factor	63/65/123
epsM	Ccl Cj	putative transferase	0/19/23
ubiE_3	Ccl Cf Cin Cj Cl Csu	SAM-dependent methyltransferase	48/48/48
cj1421c	Cin Cj Csu	putative sugar transferase	45/52/55
cj1426c	Cj	putative methyltransferase family protein	2/8/8
cj1429c	Cj Csu	hypothetical protein Cj1429c	14/14/14
cj1437c	Cj	aminotransferase	7/7/8
cJJ81176_0082	Cj	conserved domain protein	0/1/45
cJJ81176_0086	Cj	anion transporter	1/1/1
cJJ81176_0590	Cj	FIG00470712: hypothetical protein	0/11/14
cJJ81176_0758	Cj	Putative periplasmic protein	9/9/26

Name	In species	Putative Function	# in group[1]
cJJ81176_1160	Ccl Cin Cj Csu	beta-1,4-N-acetylgalactosaminyltransferase	11/11/39
lgrA	Ccl Chy Cj Cl Cv	formyl transferase domain protein	21/21/29
hyaD_3	Ccl Cj	galactosyltransferase	7/7/22
ansA	Ccl Cj	L-asparaginase	0/24/78
sodB	Cj	superoxide dismutase (Fe)	0/3/76
kdpB	Cj	potassium-transporting ATPase subunit B	5/7/33
a911_05550	Cj	Hypothetical protein	1/1/1
a911_06917	Chy Cj	putative phospholipase / phosphodiesterase	5/7/10
a911_07000	Ccl Cj Cl Csp Csu	sugar transferase	34/34/50
pJ16_02800	Cj	membrane protein	0/8/31
pJ16_05965	Cj	hypothetical protein	1/1/1
pJ17_02095	Cj	putative periplasmic ATP /GTP-binding protein	4/5/58
era_1	Cj	GTP-binding protein	0/6/6
pJ17_03155	Cj	integral membrane protein	6/6/46
pJ17_03270	Cj	No annotation data	1/1/3
cjeI	Ccl Ccn Cf Cj	restriction endonuclease	21/21/47
cstI	Ccl Cj Csu	alpha-2,3 sialyltransferase	5/5/25
pJ17_06935	Ccl Cj Cl	3-oxoacyl-ACP synthase	10/10/11
pJ17_07195	Cj	vacuolating cytotoxin precursor	0/8/9
pJ17_07500	Cj	hypothetical protein	0/1/1
pJ17_08700	Cj	hypothetical protein	0/1/3
n135_02190	Cj	hypothetical protein	1/1/1
n564_01110	Cj	hypothetical protein	1/1/1
n565_02255	Cj	hypothetical protein	1/1/1
n755_02315	Cj	hypothetical protein	1/1/1
vatD	Cj Csu	xenobiotic acyltransferase (XAT) family acetyltransferase	3/3/15
pJ18_06805	Ccl Cin Cj Csu Cv	N-acetyl sugar amidotransferase	18/20/50
pJ18_07285	Cj	hypothetical protein	2/2/10
gmd	Cin Cj Csp	GDP-mannose 4,6-dehydratase	10/10/28
pJ18_07325	Ccl Cj	sugar transferase	5/5/7
pJ19_06230	Cj	hypothetical protein	1/1/1
jJD26997_0066	Cj	conserved hypothetical protein	1/1/1
jJD26997_0382	Cj	deacetylase, PIG-L family	2/2/4
jJD26997_0387	Cj	No annotation data	1/1/1

Name	In species	Putative Function	# in group[1]
jJD26997_0403	Cj	No annotation data	1/1/12
jJD26997_0405	Cj Cl Csu	methyltransferase	8/8/23
jJD26997_0569	Cj	copper-translocating P-type ATPase	0/1/58
jJD26997_0585	Cj	beta-1,3-galactosyltransferase	1/1/7
pglF	Cj	No annotation data	1/1/77
jJD26997_0628	Cj	No annotation data	1/2/2
jJD26997_0684	Cj Csu	polysaccharide deacetylase family protein	4/4/4
jJD26997_0701	Cj	No annotation data	1/1/1
fcl	Cj Csp Csu	GDP-fucose synthetase	12/12/48
jJD26997_0710	Cj	WbdK	1/1/9
jJD26997_0711	Cj	glycosyl transferase family 8	1/1/1
jJD26997_0715	Cj	galactosyltransferase	1/1/1
jJD26997_0718	Cj	putative sugar transferase	1/1/1
jJD26997_0721	Cj	conserved hypothetical protein	1/2/2
jJD26997_0722	Ccl Cj	putative acylneuraminate cytidylyltransferase	3/3/4
jJD26997_0724	Cj	hypothetical protein	1/1/1
jJD26997_0873	Cj	No annotation data	1/1/1
cheR	Cj	No annotation data	1/1/58
jJD26997_0946	Cj	conserved hypothetical protein	0/1/3
jJD26997_1036	Cj	No annotation data	0/1/2
jJD26997_1094	Cj	No annotation data	1/1/1
jJD26997_1251	Cj	methyltransferase, FkbM family	2/2/2
jJD26997_1276	Cj	hypothetical protein	1/1/1
jJD26997_1424	Cj	conserved hypothetical protein	0/1/7
jJD26997_1619	Cj	No annotation data	0/1/2
kpsS	Cj	capsule polysaccharide export protein KpsS	0/6/58
hddC	Cj	D-glycero-D-manno-heptose 1-phosphate guanosyltransferase	3/4/36
jJD26997_1758	Cj	No annotation data	1/1/1
jJD26997_1799	Cin Cj	SAM-dependent methyltransferase	6/6/12
jJD26997_1801	Cj	capsular polysaccharide biosynthesis protein	1/1/11
jJD26997_1809	Cj	putative sugar transferase	3/8/9
jJD26997_1851	Cj	putative type I restriction modification DNA specificity domain	0/1/1

Name	In species	Putative Function	# in group[1]
tsr	Cj	methyl-accepting chemotaxis signal transduction protein	0/2/184
murI	Cj	glutamate racemase	1/1/58
bN867_11290	Cj Cl	dTDP-6-deoxy-3,4-keto-hexulose isomerase	8/8/11
bN867_11310	Cj	dTDP-6-deoxy-3,4-keto-hexulose isomerase	5/5/11
vacA	Ccl Cj	autotransporter	0/12/27
bN867_13990	Cj	putative sugar transferase	3/4/4
bN867_14120	Cj	glycosyl transferase family 2	5/5/56
m635_01965	Cj	DNA glycosylase	0/5/57
m635_02500	Cj	hypothetical protein	0/1/1
#103	Cj	No data, identified by script	1/1/1
uC78_0928	Cj	hypothetical protein	1/2/43
legI	Cj	N,N'-diacetyllegionaminic acid synthase	0/2/43
uC78_1371	Cj	F5/8 type C domain protein	1/1/1
uC78_1372	Cj	hypothetical protein	1/1/1
uC78_1398	Cj	FlgN protein	1/1/59
c8J_0464	Cj	hypothetical protein	0/2/11
pseH	Cj Cl Cp Csp Csu	UDP-4-amino-4,6-dideoxy-beta-L-AltNAc o-acetyltransferase	11/11/61
c8J_1334	Cj	hypothetical protein	1/1/1
c8J_1341	Cj	hypothetical protein	1/1/1
kpsF	Cj	arabinose-5-phosphate isomerase	1/1/59
cJ8421_r08795	Cj Cl	5S ribosomal RNA	0/3/199
cJ8421_03315	Cj	potassium-transporting ATPase subunit B	0/1/3
cJ8421_03620	Cj	Histidine-binding protein	0/1/58
cJ8421_04090	Cj	bifunctional phosphopan-tothenoylcysteine decarboxy-lase/phosphopantothenate synthase	0/1/65
cJ8421_06510	Ccl Cj	Putative amino acid activating enzyme	0/2/67
cJ8421_06565	Cj	motility accessory factor	1/1/1
cJ8421_06570	Cj	motility accessory factor	1/1/1
cJ8421_07195	Cj	putative nucleotide-sugar epimerase-dehydratase	0/2/34
hemH	Cho Cj	ferrochelatase	1/2/78
rC25_03370	Cj	No annotation data	1/1/1
rC25_04535	Cj	acyl-CoA thioester hydrolase	0/1/58

Name	In species	Putative Function	# in group[1]
cJH_04645	Ccu Cj	argininosuccinate lyase	1/2/64
cJH_06685	Cj	hypothetical protein	1/1/1
cJSA_1086	Cj	hypothetical protein	1/1/1
sucC	Cj	succinyl-CoA synthetase, beta subunit	3/3/56
iCDCCJ07001_711	Cj	adhesive protein CupB5, putative	3/3/13
pseA	Ccl Cj Cl	pseudaminic acid biosynthesis protein	1/5/54
iCDCCJ07001_1264	Cj	Protein of unknown function family	2/2/3
iCDCCJ07001_1355	Cj	putative sugar transferase	3/3/3
iCDCCJ07001_1365	Cj	putative nucleotidyl-sugar epimerase	3/3/6
iCDCCJ07001_1596	Cj	major facilitator family protein	1/1/17
cJM1_0674	Cj	probable membrane protein	1/1/45
mTVDSCj20_0757	Cj	No annotation data	1/1/1
eRS445056_00970	Cj	Uncharacterised protein	1/1/1
rpsO_2	Ccl Cj	30S ribosomal protein S15	0/3/4
eRS445056_01018	Cj	Uncharacterised protein	0/1/6
eRS445056_01054	Cj	Uncharacterised protein	1/1/6
kfoC	Cj	lipooligosaccharide biosynthesis galactosyltransferase	1/1/5
eRS445056_01385	Cj	M24/M37 family peptidase	1/1/73
cstI	Cj	sialyltransferase CST-I%2C CAzY family GT42	2/2/2
eRS445056_01580	Cj	putative sugar transferase	1/1/1
eRS445056_01623	Cj	type I restriction modification DNA specificity domain-containing protein	1/1/1
a0W68_02765	Cj	hemolysin	0/1/73
a0W68_03505	Cj	AMP-dependent synthetase	0/3/9
a0W68_05955	Cj	hypothetical protein	1/1/1
h730_00350	Cj	3-dehydroquinate dehydratase	0/1/77
h730_00775	Cj	hypothetical protein	0/1/1
h730_04015	Cj	hydrogenase nickel insertion protein HypA	0/1/57
h730_04210	Cj	hypothetical protein	1/1/77
h730_06665	Cj	lipooligosaccharide biosynthesis galactosyltransferase	3/3/9
h730_07480	Cj	hypothetical protein	0/1/34
h730_07535	Ccl Cj	N-acetylneuraminic acid synthetase	0/2/63

Name	In species	Putative Function	# in group[1]
h730_07555	Cj	DegT family aminotransferase	0/1/37
cJE1105	Cj	hypothetical protein	1/1/2
cJE1515	Cj	formyltransferase, putative	2/2/2
cJE1602	Cj	Capsular polysaccharide biosynthesis heptosyltransferase HddD	2/2/2
cJE1603	Cj	Capsular polysaccharide biosynthesis heptosyltransferase HddD	2/2/2
cjjRM1285_1441	Cj	putative sulfotransferase	1/1/4
cjjRM1285_1443	Cj	hypothetical protein	0/1/1
cjjRM1285_1570	Cj	No annotation data	1/1/11
aXW77_00885	Cj	TonB-dependent receptor	1/1/35
a0W69_06915	Cj	hypothetical protein	1/1/1
legI_1	Cj	N,N'-diacetyllegionaminic acid synthase	0/1/28
murD	Ccl	UDP-N-acetylmuramoylalanine–D-glutamate ligase	0/9/76
n149_0842	Ccl	Hypothetical protein	0/10/13
n149_0993	Ccl Cl Cp Csp Csu Cv	Phosphoglycerol transferase	17/17/24
hxuA	Ccl	filamentous hemagglutinin	5/12/16
n149_1379	Ccl	Hypothetical protein	2/2/3
n149_1457	Ccl Cin Cl Cp Csu	autotransporter domain protein	0/10/12
aB430_02410	Ccl	hypothetical protein	0/1/10
aB430_08635	Ccl Cf	Ferrous iron transport protein B	3/3/70
g157_01750	Ccl	capsular polysaccharide biosynthesis protein	1/1/1
yjjG	Ccl Cin Cl	haloacid dehalogenase-like hydrolase domain/phosphoribulokinase domain protein	5/5/10
g157_02350	Ccl	aminotransferase, DegT/DnrJ/EryC1/StrS family protein	0/1/72
g157_03665	Ccl	hypothetical protein	0/1/75
g157_07480	Ccl Chy	cysteine permease	2/3/74
vC76_07095	Ccl	hypothetical protein	1/1/4
aR446_00750	Ccl	formate dehydrogenase	0/2/77
aR446_02585	Ccl	hypothetical protein	4/4/13

Name	In species	Putative Function	# in group[1]
aR446_05305	Ccl	rod shape-determining protein RodA	2/2/78
aR446_05885	Ccl	citrate lyase	3/3/10
cysQ_1	Ccl	3′(2′),5′-bisphosphate nucleotidase CysQ	0/1/5
hxuA	Ccl	Heme/hemopexin-binding protein precursor	1/1/1
pepD_2	Ccl	Cytosol non-specific dipeptidase	1/1/58
aTE51_02562	Ccl	hypothetical protein	0/2/2
ySS_00300	Ccl	No annotation data	1/1/38
ySS_01045	Ccl	No annotation data	0/1/1
ySS_03440	Ccl	hypothetical protein	1/1/15
ySS_04555	Ccl	hypothetical protein	0/1/3
ySS_09825	Ccl	flagellar hook protein FlgE	1/1/58
cCC13826_1135	Ccn	hypothetical protein	0/1/1
miaA	Ccn	tRNA(i6A37) synthase	0/1/19
cCC13826_1936	Ccn	hypothetical protein	1/1/1
pykF	Ccn	pyruvate kinase I	0/1/77
cCC13826_1603	Ccn	hypothetical protein	0/1/1
cCC13826_0878	Ccn	hypothetical protein	1/1/1
cCC13826_1611	Ccn	Na+-dependent transporter, SNF family	0/1/77
cCC13826_1097	Ccn	hypothetical protein (GDYXXLXY domain)	1/1/1
cCC13826_0127	Ccn	hypothetical protein	0/1/2
cCC13826_0604	Ccn	hypothetical protein	0/1/2
cCC13826_2027	Ccn	hypothetical protein	0/1/1
cCC13826_0410	Ccn	zinc-dependent peptidase, M16 family	0/1/3
cCC13826_0523	Ccn	glycosyltransferase, family 1	1/1/1
asnB2	Ccn	asparagine synthase (glutamine-hydrolyzing)	1/1/2
cCC13826_0538	Ccn	hypothetical protein	1/1/2
cCC13826_1864	Ccn	hypothetical protein	1/1/1
cCON33237_0131	Ccn	No annotation data	1/1/1
accD	Ccn	acetyl-CoA carboxylase, carboxyltransferase, beta subunit	0/1/77
cCON33237_0535	Ccn	No annotation data	1/1/1
cCON33237_1104	Ccn	terminase domain protein	0/1/11
cCON33237_1186	Ccn	metallophosphatase	0/1/1
cCON33237_1757	Ccn	type II restriction endonuclease	3/3/3
cCV52592_1110	Ccu	TonB-dependent heme/hemoglobin receptor family protein	1/1/1

Name	In species	Putative Function	# in group[1]
cCV52592_0842	Ccu	hypothetical protein	0/1/1
cCV52592_0871	Ccu	No annotation data	1/1/1
cCV52592_1239	Ccu	TonB-dependent receptor	0/1/4
cCV52592_1369	Ccu Chy	outer membrane beta-barrel domain protein	0/2/11
cCV52592_1407	Ccu	competence protein ComEA helix-hairpin-helix repeat protein	0/1/1
cCV52592_1309	Ccu	hypothetical protein	0/1/2
cCV52592_2185	Ccu	No annotation data	0/1/146
cCV52592_2212	Ccu	TonB-dependent receptor	3/3/3
pncBII	Ccu	nicotinate phosphoribosyltransferase, subgroup B	0/1/3
cFT03427_0448	Cf	hypothetical protein	0/3/8
sapA3	Cf	surface array protein A	0/1/17
cFT03427_0509	Cf	hypothetical protein (DUF945 domain)	0/3/8
cFT03427_0515	Cf	MCP-domain signal transduction protein (chemoreceptor zinc-binding domain)	0/3/11
menA	Cf Chy	1,4-dihydroxy-2-naphthoate octaprenyltransferase	0/8/77
cFT03427_0876	Cf	SAM-dependent methyltransferase	8/8/8
cFT03427_0951	Cf	hypothetical protein	16/16/16
cFT03427_0957	Cf Chy	type III restriction/modification system, mod subunit	4/4/6
cFT03427_1021	Cf	MCP-domain signal transduction protein	6/8/48
cFT03427_1099	Cf	putative membrane protein	8/8/8
cFT03427_1115	Cf Chy	autotransporter domain protein	19/20/21
cFT03427_1417	Cf	hypothetical protein	1/1/8
cFT03427_1442	Cf	transformation system protein	6/7/9
cFT03427_1510	Cf Chy	ATP-grasp domain protein	12/12/12
cFT03427_1510	Cf	ATP-grasp domain protein	1/1/1
cFT03427_1512	Cf Chy	SAM-dependent methyltransferase	9/9/15
cFT03427_1545	Cf	radical SAM superfamily enzyme, MoaA/NifB/PqqE/SkfB family	7/7/8

Name	In species	Putative Function	# in group[1]
cFT03427_1551	Cf	short-chain dehydrogenase/reductase family protein	7/7/7
cFT03427_1554	Cf	methyltransferase	7/7/8
cFT03427_1554	Cf	methyltransferase	1/1/1
cFT03427_1556	Cf	formyltransferase domain-containing protein	6/6/6
cFT03427_1558	Cf	radical SAM superfamily enzyme, MoaA/NifB/PqqE/SkfB family (SPASM domain)	7/7/14
cFT03427_1559	Cf	hypothetical protein	7/7/8
cFT03427_1562	Cf	probable 3-demethylubiquinone-9 3-methyltransferase	8/8/8
cFT03427_1565	Cf	hypothetical protein	7/7/8
cFT03427_1566	Cf	hypothetical protein	7/7/8
cFT03427_1573	Cf	hypothetical protein	8/8/8
cFT03427_1574	Cf	hypothetical membrane protein	7/8/8
cFT03427_1577	Cf	hypothetical membrane protein	7/7/7
cFT03427_1581	Cf	SAM-dependent methyltransferase	8/8/48
cFT03427_1684	Cf	hypothetical protein	8/8/8
cFT03427_1756	Cf	hypothetical protein	0/3/3
cFF04554_0485	Cf	glycosyltransferase, family 1	1/1/1
cFF04554_0871	Cf	putative type II secretion system protein	5/5/8
crcB	Cf	putative fluoride ion transporter	0/1/40
cFF04554_1097	Cf	No annotation data	1/1/9
cFF04554_1106	Cf	hypothetical protein	0/4/4
cFF04554_1255	Cf	4HB_MCP sensor-containing MCP-domain signal transduction protein	5/5/8
cFF04554_1562	Cf	ATP-grasp domain-containing protein	1/1/1
cFF04554_1600	Cf	radical SAM superfamily enzyme, MoaA/NifB/PqqE/SkfB family	2/2/9
cFF04554_1615	Cf	hypothetical protein	1/1/1
cFF04554_1707	Cf	hypothetical protein	0/2/7
cFF8240_1594	Cf	hypothetical protein	1/1/1
cFF8240_1596	Cf	UDP-glucose 4-epimerase, putative	1/1/1

Name	In species	Putative Function	# in group[1]
cFF8240_1601	Cf	heme biosynthesis protein, putative	1/1/1
cFF8240_1603	Cf	methyltransferase	1/1/1
cFF8240_1604	Cf	hypothetical protein	1/1/1
cFF8240_1614	Cf	hypothetical protein	1/1/1
cSG_5300	Cf	Putative periplasmic protein	2/3/7
cSG_6080	Cf	hypothetical protein	2/2/2
cSG_7140	Cf	hypothetical protein	0/2/6
cSG_13710	Cf	type IV secretion system protein VirB2	2/3/3
cSG_18950	Cf Chy	hypothetical protein	0/4/4
cFV97608_1684	Cf	No annotation data	1/1/1
sapA3	Cf	surface array protein A	0/1/9
sapA5	Cf	surface array protein A	0/1/23
cFTSP3_1574	Cf	No annotation data	1/1/1
cFTSP3_1618	Cf	methyltransferase	1/1/1
cFVI03293_0814	Cf	No annotation data	1/1/76
cFVI03293_1049	Cf	hypothetical protein	1/1/1
cFVI03293_1574	Cf	No annotation data	1/1/1
cR44_02375	Cf	ATPase AAA	0/1/8
cR44_02485	Cf	cell surface protein	0/2/7
cR44_07880	Cf	hypothetical protein	1/1/1
cR44_07895	Cf	hypothetical protein	0/1/7
cGRAC_0030		zinc metallopeptidase, M48 family	0/1/18
cGRAC_0049		hypothetical protein	0/1/1
iscU		iron-sulfur cluster assembly scaffold protein IscU	1/1/77
cGRAC_0152		hypothetical protein	1/1/1
nuoB		NADH:quinone oxidoreductase I, chain B	1/1/64
cGRAC_0512		FAD-dependent oxidoreductase	1/1/1
cGRAC_0634		adenine-specific DNA methyltransferase	1/1/38
cGRAC_0654		hypothetical protein	0/1/1
cGRAC_1031		hypothetical protein	0/1/1
cGRAC_1153		No annotation data	1/1/1
msrAB		bifunctional (RS)-methionine sulfoxide reductase A/B	0/1/19
cGRAC_1463		hypothetical protein	0/1/1
cGRAC_1682		hypothetical protein	0/1/1
cGRAC_1691		hypothetical protein	0/1/67
cGRAC_1754		hypothetical protein	0/1/1
polA		DNA polymerase I, 5′ –> 3′ polymerase, 5′ –> 3′ and 3′ –> 5′ exonuclease	1/1/77

Name	In species	Putative Function	# in group[1]
cGRAC_1916		acyltransferase	1/1/1
cGRAC_1933		Sel1 domain protein	0/1/10
cGRAC_1935		hypothetical protein	1/1/1
cGRAC_2035		hypothetical protein	0/1/2
cHAB381_0135	Cho Cu	site-specific recombinase, phage integrase family	0/2/6
cHAB381_0570	Cho	hypothetical protein	1/1/1
cHAB381_0791	Cho	No annotation data	1/1/1
cHAB381_0867	Cho	conserved hypothetical protein	0/1/1
cHAB381_1125	Cho	conserved hypothetical protein	0/1/1
cHAB381_1453	Cho	NAD-dependent deacetylase (Regulatory protein SIR2-like protein)	0/1/55
cHAB381_1470	Cho	hypothetical protein	1/1/1
cHAB381_1478	Cho	NDP-N-acetyl-D-galactosaminuronic acid dehydrogenase	1/1/6
cHAB381_1579	Cho	hypothetical protein	0/1/1
cHAB381_1727	Cho	hypothetical protein	1/1/1
cHL_0019	Chy	phosphoglycerol transferase	5/5/7
cHL_0019	Chy	No annotation data	0/1/1
cHL_0034	Chy	PAS sensor-containing diguanylate cyclase/phosphodiesterase	1/1/13
cHL_0067	Chy	No annotation data	1/1/1
cHL_0068	Chy	No annotation data	1/1/1
cHL_0070	Chy	SAM-dependent methyltransferase	1/1/1
cHL_0074	Chy	metallophosphatase	1/1/2
cHL_0091	Chy	No annotation data	1/1/1
cHL_0105	Chy	type III restriction/modification system, mod subunit	1/1/1
cHL_0142	Chy	formyltransferase domain-containing protein	1/1/1
cHL_0216	Chy	hypothetical protein	0/1/1
cHL_0231	Chy	ATP-grasp domain-containing protein	5/5/9
cHL_0238	Chy	WbqC family protein	2/2/2
cHL_0240	Chy	aminotransferase, DegT/DnrJ/EryC1/StrS family	5/5/5
cHL_0243	Chy	No annotation data	1/1/1
cHL_0291	Chy	hypothetical protein	0/1/1
cHL_0424	Chy	TonB-dependent receptor	2/2/2

Name	In species	Putative Function	# in group[1]
cHL_0448	Chy	SAM-dependent methyltransferase	1/1/1
cHL_0502	Chy	glycosyltransferase, family 1	2/2/5
cmoA	Chy	carboxy-S-adenosyl-L-methionine synthase	2/2/77
cHL_0706	Chy	diguanylate cyclase	2/2/3
cHL_0780	Chy	autotransporter domain protein	7/10/10
cHL_0780	Chy	No annotation data	0/1/1
cHL_0782	Chy	No annotation data	0/1/1
cHL_0783	Chy	No annotation data	0/1/1
cHL_0793	Chy	hypothetical protein	0/1/1
cHL_0832	Chy	hypothetical protein	0/2/2
cHL_0862	Chy Cin Cv	membrane bound O-acyl transferase, MBOAT family	7/7/25
cHL_0867	Chy	PAS sensor-containing signal-transduction protein	1/1/13
cHL_1002	Chy	hypothetical protein	0/1/10
cHL_1115	Chy	No annotation data	1/1/1
cHL_1160	Chy Csp	glycosyltransferase, family 8	2/2/2
cHL_1240	Chy	hypothetical protein	0/1/6
cHL_1325	Chy	receiver domain protein	0/2/13
cHL_1435	Chy	aminoglycoside N3'-acetyltransferase	1/1/1
cHL_1437	Chy	short-chain dehydrogenase/reductase	1/1/1
fla1	Chy	flagellin	1/2/27
cHL_1682	Chy	hypothetical protein	0/1/1
hsdS1	Chy Cl	type I restriction/modification system, S subunit	3/3/3
cHL_1707	Chy	No annotation data	0/1/1
cHH_0008	Chy	IS605/IS607 family integrase/resolvase	0/1/18
cHH_0019	Chy	putative membrane protein, EpsG family	1/4/4
cHH_0020	Chy	glycosyltransferase, family 1	3/3/6
cHH_0023	Chy	putative serine acetyltransferase	2/2/2
cHH_0024	Chy	methyltransferase	2/2/2
cHH_0029	Chy	putative Zn-peptidase, M28 family (DUF2172 domain)	2/2/2
cHH_0032	Chy	transferase hexapeptide repeat containing protein, putative	2/2/2
cHH_0033	Chy	methyltransferase	1/1/1

Name	In species	Putative Function	# in group[1]
cHH_0038	Chy	methyltransferase FkbM family protein, putative	2/2/2
cHH_0041	Chy	transferase hexapeptide repeat containing protein, putative	2/2/3
cHH_0042	Chy	xenobiotic acyltransferase (XAT) family acetyltransferase	1/1/2
cHH_0054	Chy	SAM-dependent methyltransferase	1/1/1
dcuB	Chy	anaerobic C4-dicarboxylate transporter	0/1/74
cHH_0173	Chy	Cache sensor-containing MCP-domain signal transduction protein	0/2/41
cHH_0314	Chy	No annotation data	1/1/1
cHH_0443	Chy	hypothetical protein	1/1/1
cHH_0490	Chy	MCP-domain signal transduction protein (DUF3365 domain)	0/2/13
cHH_0528	Chy	putative ComE family competence protein	0/2/18
cHH_0538	Chy	diguanylate cyclase	0/1/12
cHH_0803	Chy	No annotation data	1/1/1
cHH_0814	Chy	outer membrane beta-barrel domain protein	0/3/3
cHH_0890	Chy	No annotation data	1/1/1
cHH_0918	Chy	No annotation data	0/1/1
cHH_1201	Chy	putative DUF945 domain protein	0/1/1
flgS	Chy	two-component system, sensor histidine kinase	1/1/53
cHH_1249	Chy	hypothetical protein	1/1/1
cHH_1302	Chy	No annotation data	0/1/1
cHH_1361	Chy	No annotation data	1/1/1
cHH_1397	Chy	autotransporter domain protein	1/1/1
cHH_1665	Chy	hypothetical protein	0/1/1
cHH_1730	Chy	MCP-domain signal transduction protein	0/1/1
cHH_1765	Chy	hypothetical protein	0/1/1
cHH_1768	Chy	hypothetical protein	1/1/1
cIG1485E_0027		sugar O-acyltransferase, sialic acid O-acetyltransferase NeuD family	1/1/2
cIG1485E_0031		formyltransferase domain-containing protein	1/1/3

Name	In species	Putative Function	# in group[1]
cIG1485E_0039		glycosyl amidation-associated protein WbuZ	1/1/1
cIG1485E_0042		SAM-dependent methyltransferase	1/1/1
cIG1485E_0045		CMP-N-acetylneuraminic acid synthetase	1/1/2
cIG1485E_0050		hypothetical protein	1/1/1
cIG1485E_0051		No annotation data	1/1/1
cIG1485E_0052		SAM-dependent methyltransferase	1/1/1
cIG1485E_0074		hypothetical type II secretion system protein	2/2/3
cIG1485E_0147		glucose-1-phosphate thymidylyltransferase	0/1/16
cIG1485E_0289		TonB-dependent receptor	3/3/11
cIG1485E_0543		hypothetical protein	0/2/2
cIG1485E_0617		autotransporter domain protein	2/2/2
cIG1485E_0654		autotransporter domain protein	2/2/2
cIG1485E_0808		No annotation data	1/1/1
cIG1485E_1074		hypothetical protein	0/2/2
cIG1485E_1167		hypothetical membrane protein	3/3/3
cIG1485E_1224		hypothetical protein	0/1/1
cIG1485E_1451		hypothetical protein	0/2/5
cIG2463D_0029		hypothetical protein	1/1/1
cIG2463D_0038		polysaccharide biosynthesis protein, putative	1/1/2
cIG2463D_0043		NAD dependent epimerase/dehydratase, putative	1/1/2
cIG2463D_0045		methyltransferase, putative	1/1/1
cIG2463D_0046		No annotation data	1/1/1
cIG2463D_0055		epimerase/dehydratase, putative	1/1/2
cIG2463D_0063		hypothetical protein	1/1/1
cIG2463D_0120		No annotation data	1/1/11
cIG2463D_0809		No annotation data	1/1/1
cIG2463D_0979		regulator of chromosome condensation RCC1, putative	2/2/2
cIG2463D_0981		phage tail fiber protein, putative	1/1/2
cIG2463D_1478		hypothetical protein	1/1/2
cIG2463D_1832		family 2 glycosyl transferase, putative	1/1/1
cIG11343_0027		glycosyltransferase, family 2	0/1/1

Name	In species	Putative Function	# in group[1]
cIG11343_0071		formyltransferase domain-containing protein	1/1/1
cIG11343_0143		No annotation data	0/1/1
cIG11343_0309		cytosine-specific DNA methyltransferase	1/1/1
cIG11343_0596		hypothetical protein	0/1/1
cIG11343_0600		No annotation data	1/1/1
cIG11343_0760		No annotation data	1/1/1
cIG11343_1047		hypothetical protein	1/1/1
cIG11343_1355		glycosyltransferase, family 1	1/1/5
cIG11343_1452		SAM-dependent methyltransferase	1/1/7
cIG11343_1618		cytosine-specific DNA methyltransferase	1/1/4
por	Cin	pyruvate:ferredoxin (flavodoxin) oxidoreductase, homodimeric	0/1/78
modB	Cin Cl	molybdenum ABC transporter ModABC, permease protein	0/2/32
cINS_1208	Cin	glycosyltransferase, family 2	1/1/3
cINS_1243	Cin	RND superfamily exporter	1/1/58
uPTC4110_0426	Cl	molybdopterin-containing oxidoreductase I, DMSO/TMAO/BSO reductase family, catalytic subunit	0/3/28
uPTC4110_0579	Cl Csu Cv	autotransporter domain protein	6/7/7
uPTC4110_0710	Cl Cp Csp Csu Cv	No annotation data	30/31/45
uPTC4110_0958	Cl Csu	autotransporter domain protein	0/6/7
uPTC4110_1065	Cl Csp	hypothetical protein (DUF2920 domain)	3/3/6
uPTC4110_1471	Cl Cp Csp	MCP-domain signal transduction protein	0/7/11
uPTC4110_1503	Cl	hemagglutinin domain-containing protein	1/1/2
cONCH_0029	Cl	No annotation data	1/1/1
cONCH_0134	Cl	hypothetical protein	1/1/1
cONCH_0312	Cl	hypothetical protein, putative serine acetyltransferase	1/1/1
cONCH_0314	Cl	glycosyltransferase, family 2	1/1/1
mqo	Cl	malate:quinone-oxidoreductase	0/1/76
cONCH_0844	Cl	DJ-1/PfpI family protein	1/1/58

Name	In species	Putative Function	# in group[1]
uPTC3659_0047	Cl Csu	GTP binding protein	0/2/2
uPTC3659_0069	Cl	autotransporter domain protein	1/2/2
uPTC3659_0279	Cl	hypothetical protein (DUF2972 domain)	1/1/1
uPTC3659_0318	Cl	glycosyltransferase, family 1	1/1/2
uPTC3659_0318	Cl	glycosyltransferase, family 1	1/1/1
uPTC3659_0363	Cl	No annotation data	1/1/1
lysS	Cl	lysyl-tRNA synthetase	1/1/77
uPTC3659_0805	Cl	MiaB-like tRNA modifying enzyme	1/1/77
uPTC3659_0845	Cl	No annotation data	1/1/1
uPTC3659_1146	Cl Cp Csp	hemagglutinin domain-containing protein	16/17/44
uPTC3659_1415	Cl Csp	sulfatase family protein	2/2/27
opgE	Cl	phosphoethanolamine transferase	1/1/1
uPTC3659_1538	Cl	hypothetical protein	1/1/2
cla_0310	Cl Csu	putative capsular polysaccharide biosynthesis protein	3/3/7
cla_0795	Cl	conserved hypothetical integral membrane protein, MATE family efflux protein	0/1/64
cla_1236	Cl	formyltransferase, putative	1/1/1
cla_1238	Cl	No annotation data	1/1/1
cla_1238	Cl	No annotation data	1/1/1
uPTC16701_0039	Cl Csu	hypothetical protein	0/2/2
uPTC16701_0833	Cl	putative adhesin/invasin (DUF3442 domain)	0/1/2
uPTC16701_1207	Cl	glucose-1-phosphate cytidylyltransferase	1/1/18
uPTC16701_1426	Cl	putative DNA helicase (AAA domain)	0/1/5
uPTC16712_0018	Cl Csu	autotransporter domain protein	6/6/7
uPTC16712_0041	Cl	hypothetical protein	0/1/6
uPTC16712_0062	Cl	autotransporter domain protein	1/1/1
uPTC16712_0309	Cl	putative capsular polysaccharide biosynthesis protein	1/1/1
cD56_03390	Cl	hypothetical protein	0/1/5
cD56_03805	Cl	MFS transporter	0/1/58
cD56_05085	Cl	aminobenzoate synthetase	0/1/56
cD56_06010	Cl	hypothetical protein	0/1/24
cPEL_0032	Cp	GTPase family protein	0/1/1

Name	In species	Putative Function	# in group[1]
cPEL_0301	Cp	hypothetical protein (DUF2172 domain), putative M28 family zinc peptidase	1/1/1
opgE	Cp Csu	phosphoethanolamine transferase	2/2/4
cAQ16704_0016	Csp	putative transporter	1/1/2
cAQ16704_0022	Csp	putative glycosyltransferase, family 9	0/1/4
flaA	Csp	flagellin	0/1/142
cAQ16704_0310	Csp	hypothetical protein	1/1/1
dsbA2	Csp	protein disulfide oxidoreductase	1/1/42
cAQ16704_0605	Csp	hypothetical protein	1/1/1
cAQ16704_0678	Csp	hypothetical protein, putative family 8 glycosyltransferase	1/1/13
cAQ16704_0713	Csp	No annotation data	0/1/9
cAQ16704_0789	Csp	putative LPT lipopolysaccharide transport system ATP-binding component LptB	1/1/77
cAQ16704_0793	Csp Csu	hypothetical protein	0/2/3
hydD	Csp	[NiFe] hydrogenase maturation protease HydD	1/1/70
cAQ16704_0868	Csp	hypothetical protein	0/2/11
cAQ16704_0969	Csp	hypothetical protein	0/1/1
cAQ16704_0978	Csp	flavocytochrome c, flavin subunit	0/1/4
cAQ16704_1019	Csp	MCP-domain signal transduction protein	0/1/58
cAQ16704_1067	Csp	hypothetical membrane protein (DUF2165 domain)	0/1/28
ccsB	Csp	cytochrome c biogenesis protein	1/1/57
cAQ16704_1161	Csp	hemerythrin-like metal-binding domain protein	0/1/7
cAQ16704_1250	Csp	glycosyltransferase, family 1	1/1/9
cAQ16704_1312	Csp	hypothetical protein (DUF2172 domain), putative M28 family zinc peptidase	1/1/1
cAQ16704_1321	Csp	short chain dehydrogenase/reductase family oxidoreductase	1/1/11
cAQ16704_1323	Csp	hypothetical protein	1/1/1
dnaG	Csp	DNA primase	1/1/59
cSUB8521_0274	Csu	No annotation data	1/1/77
argH	Csu	argininosuccinate lyase	1/1/13

Name	In species	Putative Function	# in group[1]
cSUB8521_0318	Csu	hypothetical protein	1/1/6
cSUB8521_0371	Csu	No annotation data	2/2/42
racS	Csu	two-component system sensor histidine kinase	1/1/13
cSUB8521_0608	Csu	No annotation data	1/1/1
cSUB8521_1429	Csu	glycosyltransferase, family 2	1/1/2
cSUB8521_1458	Csu	type III restriction/modification enzyme, mod subunit	2/2/5
cSUB8521_1489	Csu	hypothetical protein (DUF2172 domain), putative M28 family zinc peptidase	1/1/1
cSUB8521_1494	Csu Cv	hypothetical protein, possible glyoxalase/bleomycin resistance protein	3/3/36
cSUB8521_1658	Csu	No annotation data	0/1/4
cSUB8521_1660	Csu	hypothetical protein	0/1/17
cSUB8521_1708	Csu	mechanosensitive ion channel family protein	1/1/58
cSUB8523_0312	Csu	capsular polysaccharide biosynthesis protein, putative	1/1/4
cSUB8523_0480	Csu	hypothetical protein	0/1/3
cSUB8523_0595	Csu	No annotation data	1/1/1
cSUB8523_0660	Csu	hypothetical protein	1/1/1
cysE	Csu	serine O-acetyltransferase	0/1/74
cSUB8523_1585	Csu	hypothetical protein (DUF2172 domain), putative M28 family zinc peptidase	1/1/1
cSUB8523_1754	Csu	No annotation data	0/1/1
cSUB8523_1754	Csu	No annotation data	0/1/1
cUREO_0334	Cu	type I restriction-modification system, S subunit	1/1/1
selD	Cu	selenophosphate synthetase	1/1/50
cUREO_0913	Cu	anaerobic C4-dicarboxylate transporter	0/1/76
cVOL_0259	Cv	hypothetical protein	1/1/1
cVOL_0295	Cv	glycosyltransferase, family A (ATP grasp domain)	1/1/2

[1] Three numbers are, in order, (1) total number of genes located within an ORF, (2) total number of SSRs int he group, and (3) total number of genes, including non-PV homologues.

Appendix C

Strains used in *Campylobacter* phasome analysis

List of strains used in *Campylobacter* phasome analysis, numbering refers to that used in figure 3.9.

Number	Species	Strain
1	*C. jejuni*	NCTC 11168
2	*C. jejuni*	81-176
3	*C. jejuni*	PT14
4	*C. jejuni*	00-0949
5	*C. jejuni*	00-1597
6	*C. jejuni*	00-2425
7	*C. jejuni*	00-2426
8	*C. jejuni*	00-2538
9	*C. jejuni*	00-2544
10	*C. jejuni*	00-6200
11	*C. jejuni*	01-1512
12	*C. jejuni*	269.97
13	*C. jejuni*	4031
14	*C. jejuni*	32488
15	*C. jejuni*	35925B2
16	*C. jejuni*	81116; NCTC 11828
17	*C. jejuni*	CG8421
18	*C. jejuni*	CJ677CC519
19	*C. jejuni*	CJM1cam
20	*C. jejuni*	F38011
21	*C. jejuni*	IA3902
22	*C. jejuni*	ICDCCJ07001
23	*C. jejuni*	M1
24	*C. jejuni*	MTVDSCj20
25	*C. jejuni*	NCTC11351
26	*C. jejuni*	OD267
27	*C. jejuni*	R14
28	*C. jejuni*	RM1221
29	*C. jejuni*	RM1285
30	*C. jejuni*	RM3194
31	*C. jejuni*	RM3196
32	*C. jejuni*	RM3197
33	*C. jejuni*	S3

Number	Species	Strain
34	*C. jejuni*	WP2202
35	*C. jejuni*	YH001
36	*C. coli*	15-537360
37	*C. coli*	BFR-CA-9557
38	*C. coli*	CVM N29710
39	*C. coli*	FB1
40	*C. coli*	HC2-48
41	*C. coli*	OR12
42	*C. coli*	RM1875
43	*C. coli*	RM4661
44	*C. coli*	RM5611
45	*C. coli*	YH501
46	*C. concisus*	13826
47	*C. concisus*	ATCC 33237
48	*C. curvus*	525.92
49	*C. fetus*	03-427
50	*C. fetus*	04/554
51	*C. fetus*	82-40
52	*C. fetus*	84-112
53	*C. fetus*	97/608
54	*C. fetus*	SP3
55	*C. fetus*	cfvi03/293
56	*C. fetus*	pet-3
57	*C. gracilis*	ATCC 33236
58	*C. hominis*	ATCC BAA-381
59	*C. hyointestinalis*	CCUG 27631
60	*C. hyointestinalis*	LMG 9260
61	*C. iguaniorum*	1485E
62	*C. iguaniorum*	2463D
63	*C. iguaniorum*	RM11343
64	*C. insulaenigrae*	NCTC 12927
65	*C. lari*	CCUG 22395
66	*C. lari*	LMG 11760
67	*C. lari*	NCTC 11845
68	*C. lari*	RM2100; ATCC BAA-1060D
69	*C. lari*	RM16701
70	*C. lari*	RM16712
71	*C. lari*	Slaughter Beach
72	*C. peloridis*	LMG 23910
73	*C. sp.*	RM16704
74	*C. subantarcticus*	LMG 24374
75	*C. subantarcticus*	LMG 24377
76	*C. ureolyticus*	RIGS 9880
77	*C. volucris*	LMG 24379

Bibliography

Abbott, S. L., Waddington, M., Lindquist, D., Ware, J., Cheung, W., Ely, J. and Janda, J. M. (2005). Associated with Sporadic Episodes of Bloody Gastroenteritis and Brainerd's Diarrhea, *Journal of Clinical Microbiology* **43**(2): 585–588.

Acar, M., Mettetal, J. T. and Oudenaarden, A. V. (2008). Stochastic switching as a survival strategy in fluctuating environments, *Nature Genetics* **40**(4): 471–475.

Ackermann, H. W. (2003). Bacteriophage observations and evolution, *Research in Microbiology* **154**(4): 245–251.

Ackermann, M. and Chao, L. (2006). DNA sequences shaped by selection for stability, *PLoS Genetics* **2**(2): 224–230.

Agresti, A. and Gottard, A. (2005). Comment: Randomized Confidence Intervals and the Mid-P Approach, *Statistical Science* **20**(4): 367–371.

Aidley, J. and Bayliss, C. D. (2014). Repetitive DNA: A Major Source of Genetic Diversity in Campylobacter Populations?, *in* S. K. Sheppard (ed.), *Campylobacter Ecology and Evolution*, Caister Academic Press, Poole (UK), chapter 6, pp. 55–74.

Alamro, M., Bidmos, F. A., Chan, H., Oldfield, N. J., Newton, E., Bai, X., Aidley, J., Care, R., Mattick, C., Turner, D. P. J., Neal, K. R., Ala'aldeen, D. A. A., Feavers, I., Borrow, R. and Bayliss, C. D. (2014). Phase variation mediates reductions in expression of surface proteins during persistent meningococcal carriage, *Infection and Immunity* **82**(6): 2472–84.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipmanl, D. J. (1990). Basic Local Alignment Search Tool, *Journal of Molecular Biology* **215**(3): 403–10.

Andorf, C., Dobbs, D. and Honavar, V. (2007). Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach, *BMC Bioinformatics* **8**: 284.

Andrewes, F. W. (1922). Studies in Group-Agglutination. I: The salmonella group and its antigenic structure, *Journal of Pathology and Bacteriology* **25**(4): 505–522.

Andrewes, F. W. (1925). Studies in Group-Agglutination. II.—The absorption of agglutinin in the diphasic salmonellas, *The Journal of Pathology and Bacteriology* **28**(2): 345–359.

Anjum, A., Brathwaite, K. J., Aidley, J., Connerton, P. L., Cummings, N. J., Parkhill, J., Connerton, I. and Bayliss, C. D. (2016). Phase variation of a Type IIG restriction-modification enzyme alters site-specific methylation patterns and gene expression in Campylobacter jejuni strain NCTC11168, *Nucleic Acids Research* pp. 4581–4594.

Artymovich, K., Kim, J.-S., Linz, J. E., Hall, D. F., Kelley, L. E., Kalbach, H. L., Kathariou, S., Gaymer, J. and Paschke, B. (2013). A "successful allele" at Campylobacter jejuni contingency locus Cj0170 regulates motility; "successful alleles" at locus Cj0045 are strongly associated with mouse colonization, *Food Microbiology* **34**(2): 425–30.

Ashgar, S. S. A., Oldfield, N. J., Wooldridge, K. G., Jones, M. A., Irving, G. J., Turner, D. P. J. and Ala'Aldeen, D. A. A. (2007). CapA, an autotransporter protein of Campylobacter jejuni, mediates association with human epithelial cells and colonization of the chicken gut, *Journal of Bacteriology* **189**(5): 1856–65.

Bacon, D. J., Alm, R. A., Hu, L., Hickey, T. E., Ewing, C. P., Batchelor, R. A., Trust, T. J. and Guerry, P. (2002). DNA sequence and mutational analyses of the pVir plasmid of Campylobacter jejuni 81-176, *Infection and Immunity* **70**(11): 6242–6250.

Baker, T. A. and Sauer, R. T. (2012). ClpXP, an ATP-powered unfolding and protein-degradation machine, *Biochimica et Biophysica Acta - Molecular Cell Research* **1823**(1): 15–28.

Baldvinsson, S. B. (2014). *Dynamics of Campylobacter jejuni and their bacteriophages*, Phd thesis, University of Copenhagen.

Barton, N. H. and Charlesworth, B. (1984). Genetic Revolutions, Founder Effects, and Speciation, *Annual Review Ecological Systematics* **15**: 133–164.

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015). Fitting Linear Mixed-Effects Models using lme4, *Journal of Statistical Software* **67**(1): 1–48.

Bayliss, C. D. (2009). Determinants of phase variation rate and the fitness implications of differing rates for bacterial pathogens and commensals, *FEMS Microbiology Reviews* **33**(3): 504–20.

Bayliss, C. D., Bidmos, F., Anjum, A., Manchev, V. T., Richards, R. L., Grossier, J.-P., Wooldridge, K. G., Ketley, J. M., Barrow, P., Jones, M. and Tretyakov, M. V. (2012). Phase variable genes of Campylobacter jejuni exhibit high mutation rates and specific mutational patterns but mutability

is not the major determinant of population structure during host colonization, *Nucleic Acids Research* **40**(13): 5876–89.

Bayliss, C. D., Callaghan, M. J. and Moxon, E. R. (2006). High allelic diversity in the methyltransferase gene of a phase variable type III restriction-modification system has implications for the fitness of haemophilus influenzae, *Nucleic Acids Research* **34**(14): 4046–4059.

Bayliss, C. D., Dixon, K. M. and Moxon, R. (2004). Simple sequence repeats (microsatellites): mutational mechanisms and contributions to bacterial pathogenesis. A meeting review, *FEMS Immunology & Medical Microbiology* **40**(1): 11–19.

Bayliss, C. D., Field, D. and Moxon, E. R. (2001). The simple sequence contingency loci of Haemophilus influenzae and Neisseria meningitidis, *Journal of Clinical Investigation* **107**(6): 657–662.

Bayliss, C. D. and Palmer, M. E. (2012). Evolution of simple sequence repeat-mediated phase variation in bacterial genomes, *Annals of the New York Academy of Sciences* **1267**: 39–44.

Beaumont, H. J. E., Gallie, J., Kost, C., Ferguson, G. C. and Rainey, P. B. (2009). Experimental evolution of bet hedging, *Nature* **462**(7269): 90–3.

Beier, D. and Gross, R. (2006). Regulation of bacterial virulence by two-component systems, *Current Opinion in Microbiology* **9**(2): 143–152.

Benjamin, J., Leaper, S., Owen, R. J. and Skirrow, M. B. (1983). Description of Campylobacter laridis, a New Species Comprising the Nalidixic Acid Resistant Thermophilic Carnpylobacter (NARTC) Group, *Current Microbiology* **8**: 231–238.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. (2005). GenBank, *Nucleic Acids Research* **33**(DATABASE ISS.): 34–38.

Bensted, H. J. (1925). Experiments in Agglutination and Absorption with the Salmonella Group of Organisms, *Journal of the Royal Army Medical Corps* **44**(1): 11–22.

Brás, A. M., Chatterjee, S., Wren, B. W., Newell, D. G. and Ketley, J. M. (1999). A novel campylobacter jejuni two-component regulatory system important for temperature-dependent growth and colonization, *Journal of Bacteriology* **181**(10): 3298–3302.

Brenner, S. E. (1999). Errors in genome annotation, *Trends in Genetics* **15**(4): 132–133.

Brocchieri, L. and Karlin, S. (2005). Protein length in eukaryotic and prokaryotic proteomes, *Nucleic Acids Research* **33**(10): 3390–3400.

Burgos-Portugal, J. A., Kaakoush, N. O., Raftery, M. J. and Mitchell, H. M. (2012). Pathogenic potential of Campylobacter ureolyticus, *Infection and Immunity* **80**(2): 883–890.

Butzler, J. P., Dekeyser, P., Detrain, M. and Dehaen, F. (1973). Related vibrio in stools, *The Journal of Pediatrics* **82**(3): 493–495.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. L. (2009). BLAST plus: architecture and applications, *BMC Bioinformatics* **10**(421): 1.

Carrillo, C. D., Taboada, E., Nash, J. H. E., Lanthier, P., Kelly, J., Lau, P. C., Verhulp, R., Mykytczuk, O., Sy, J., Findlay, W. A., Amoako, K., Gomis, S., Willson, P., Austin, J. W., Potter, A., Babiuk, L., Allan, B. and Szymanski, C. M. (2004). Genome-wide Expression Analyses of Campylobacter jejuni NCTC11168 Reveals Coordinate Regulation of Motility and Virulence by flhA, *Journal of Biological Chemistry* **279**(19): 20327–20338.

Castelo, A. T., Martins, W. and Gao, G. R. (2002). TROLL–tandem repeat occurrence locator, *Bioinformatics* **18**(4): 634–636.

Caugant, D. A., Tzanakaki, G. and Kriz, P. (2007). Lessons from meningococcal carriage studies, *FEMS Microbiology Reviews* **31**(1): 52–63.

Cerdeno-Tarraga, A. M., Patrick, S., Crossman, L. C., Blakely, G., Abratt, V., Lennard, N., Poxton, I., Duerden, B., Harris, B., Quail, M. A., Barron, A., Clark, L., Corton, C., Doggett, J., Holden, M. T., Larke, N., Line, A., Lord, A., Norbertczak, H., Ormond, D., Price, C., Rabbinowitsch, E., Woodward, J., Barrell, B. and Parkhill, J. (2005). Extensive DNA inversions in the B. fragilis genome control variable gene expression, *Science* **307**(5714): 1463–1465.

Champion, O. (2005). Comparative phylogenomics of the food-borne pathogen Campylobacter jejuni reveals genetic markers predictive of infection source, *Proceedings of the National Academy of Sciences of the United States of America* **102**(44): 16043–8.

Champion, O. L., Karlyshev, A. V., Senior, N. J., Woodward, M., La Ragione, R., Howard, S. L., Wren, B. W. and Titball, R. W. (2010). Insect infection model for Campylobacter jejuni reveals that O-methyl phosphoramidate has insecticidal activity, *The Journal of Infectious Diseases* **201**(5): 776–82.

Chart, H., Frost, J. A., Oza, A., Thwaites, R., Gillanders, S. and Rowe, B. (1996). Heat-stable serotyping antigens expressed by strains of Campylobacterjejuni are probably capsular and not long- chain lipopolysaccharide, *Journal of Applied Bacteriology* **81**: 635–640.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. and De Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics, *Bioinformatics* **25**(11): 1422–1423.

Coen, P. G., Heath, P. T., Barbour, M. L. and Garnett, G. P. (1998). Mathematical models of Haemophilus influenzae type b, *Epidemiology & Infection* **120**(3): 281–295.

Cohn, M. T., Ingmer, H., Mulholland, F., Jørgensen, K., Wells, J. M. and Brøndsted, L. (2007). Contribution of conserved ATP-dependent proteases of Campylobacter jejuni to stress tolerance and virulence, *Applied and Environmental Microbiology* **73**(24): 7803–13.

Cold Spring Habour Protocols (2006). SM buffer with gelatin, *Cold Spring Harbor Protocols* **2006**(1): pdb.rec466.

Colles, F. M., Jones, K., Harding, R. M. and Maiden, M. C. J. (2003). Genetic Diversity of Campylobacter jejuni Isolates from Farm Animals and the Farm Environment Genetic Diversity of Campylobacter jejuni Isolates from Farm Animals and the Farm Environment, *Applied and Environmental Microbiology* **69**(12): 7409–7413.

Colles, F. M., Jones, T. A., McCarthy, N. D., Sheppard, S. K., Cody, A. J., Dingle, K. E., Dawkins, M. S. and Maiden, M. C. J. (2008). Campylobacter infection of broiler chickens in a free-range environment, *Environmental Microbiology* **10**(8): 2042–50.

Colles, F. M. and Maiden, M. C. J. (2012). Campylobacter sequence typing databases: Applications and future prospects, *Microbiology (United Kingdom)* **158**(11): 2695–2709.

Connerton, P. L., Carrillo, C. M. L., Swift, C., Dillon, E., Rees, C. E. D., Dodd, C. E. R., Frost, J., Scott, A. and Connerton, I. F. (2004). Longitudinal study of Campylobacter jejuni bacteriophages and their hosts from broiler chickens, *Applied and Environmental Microbiology* **70**(7): 3877–3883.

Connerton, P. L., Timms, A. R. and Connerton, I. F. (2011). Campylobacter bacteriophages and bacteriophage therapy, *Journal of Applied Microbiology* **111**: 255–265.

Cooper, K. K., Cooper, M. A., Zuccolo, A. and Joens, L. A. (2013). Re-sequencing of a virulent strain of Campylobacter jejuni NCTC11168 reveals potential virulence factors, *Research in Microbiology* **164**(1): 6–11.

Coward, C., van Diemen, P. M., Conlan, A. J. K., Gog, J. R., Stevens, M. P., Jones, M. A. and Maskell, D. J. (2008). Competing isogenic Campylobacter strains exhibit variable population structures in vivo, *Applied and Environmental Microbiology* **74**(12): 3857–67.

Curtis, a. H. (1913). A Motile Curved Anaerobic Bacillus in Uterine Discharges, *Journal of Infectious Diseases* **12**(2): 165–169.

Darrell, J. H., Farrell, B. C. and Mulligan, R. A. (1967). Case of Human Vibriosis, *British Medical Journal* **2**(5547): 287–289.

Dasti, J. I., Tareen, A. M., Lugert, R., Zautner, A. E. and Groß, U. (2010). Campylobacter jejuni: A brief overview on pathogenicity-associated factors and disease-mediating mechanisms, *International Journal of Medical Microbiology* **300**(4): 205–211.

De Bolle, X., Bayliss, C. D., Field, D., van de Ven, T., Saunders, N. J., Hood, D. W. and Moxon, E. R. (2000). The length of a tetranucleotide repeat tract in Haemophilus influenzae determines the phase variation rate of a gene with homology to type III DNA methyltransferases, *Molecular Microbiology* **35**(1): 211–22.

Debruyne, L., Broman, T., Bergstro, S., On, S. L. W. and Vandamme, P. (2010). Campylobacter volucris sp. nov., isolated from black-headed gulls (Larus ridibundus), *International Journal of Systematic and Evolutionary Microbiology* (60): 1870–1875.

Debruyne, L., Broman, T., Bergström, S., Olsen, B., On, S. L. W. and Vandamme, P. (2010). Campylobacter subantarcticus sp. nov., isolated from birds in the sub-Antarctic region, *International Journal of Systematic and Evolutionary Microbiology* **60**(4): 815–819.

Debruyne, L., On, S. L. W., De Brandt, E. and Vandamme, P. (2009). Novel Campylobacter lari-like bacteria from humans and molluscs: Description of Campylobacter peloridis sp. nov., Campylobacter lari subsp. concheus subsp. nov. and Campylobacter lari subsp. lari subsp. nov, *International Journal of Systematic and Evolutionary Microbiology* **59**(5): 1126–1132.

Dekeyser, P., Gossuin-Detrain, M., Butzler, J. P. and Sternon, J. (1972). Acute enteritis due to related vibrio: First positive stool cultures, *Journal of Infectious Diseases* **125**(4): 390–392.

Des Prez, R. M., Bryan, C. S. and Colley, D. G. (1975). Function of the Classical and Alternate Pathways of Human Complement in Serum Treated with Acid and MgCl 2 -Ethylene Function of the Classical and Alternate Pathways of Human Complement in Serum Treated with Ethylene Glycol Tetraacetic Acid and MgCl2-Ethy, *Infection and Immunity* **11**(6): 1235–1243.

Devos, D. and Valencia, a. (2000). Practical limits of functional prediction, *Proteins* **41**(February): 98–107.

Devos, D. and Valencia, A. (2001). Intrinsic errors in genome annotation, *Trends in Genetics* **17**(8): 429–431.

Didelot, X. and Falush, D. (2007). Inference of bacterial microevolution using multilocus sequence data, *Genetics* **175**(3): 1251–1266.

Dingle, K. E., Colles, F. M., Ure, R., Wagenaar, J. A., Duim, B., Bolton, F. J., Fox, A. J., Wareing, D. R. A. and Maiden, M. C. J. (2002). Molecular characterization of Campylobacter jejuni clones: a basis for epidemiologic investigation, *Emerging Infectious Diseases* **8**(9): 949–55.

Dingle, K. E., Colles, F. M., Wareing, D. R. A., Ure, R., Fox, A. J., Bolton, F. E., Bootsma, H. J., Willems, R. J. L., Urwin, R. and Maiden, M. C. J. (2001). Multilocus Sequence Typing System for Campylobacter jejuni, *Journal of Clinical Microbiology* **39**(1): 14–23.

Donahue, J. P., Israel, D. A., Peek, R. M., Blaser, M. J. and Miller, G. G. (2000). Overcoming the restriction barrier to plasmid transformation of Helicobacter pylori, *Molecular Microbiology* **37**(5): 1066–1074.

Dornberger, U., Leijon, M. and Fritzsche, H. (1999). High Base Pair Opening Rates in Tracts of GC Base Pairs *, *The Journal of Biological Chemistry* **274**(11): 6957–6962.

Duckworth, D. H. (1976). Who Discovered Bacteriophage?, *Bacteriological Reviews* **40**(4): 739–802.

Dworkin, J. and Blaser, M. J. (1997). Molecular mechanisms of Campylobacter fetus surface layer protein expression, *Molecular Microbiology* **26**(3): 433–440.

EFSA (2010). Analysis of the baseline survey on the prevalence of Campylobacter in broiler batches and of Campylobacter and Salmonella on broiler carcasses in the EU, 2008 - Part A: Campylobacter and Salmonella prevalence estimates, *EFSA Journal* **8**(03): 1503.

El-Shibiny, A., Connerton, P. L. and Connerton, I. F. (2005). Enumeration and Diversity of Campylobacters and Bacteriophages Isolated during the Rearing Cycles of Free-Range and Organic Chickens, *Applied and Environmental Microbiology* **71**(3): 1259–1266.

Fitzgerald, C., Whichard, J. and Nachamkin, I. (2008). Diagnosis and Antimicrobial Susceptibility of Campylobacter Species, *in* I. Nachamkin, C. M. Szymanski and M. J. Blaser (eds), *Campylobacter, 3rd edition*, ASM Press, pp. 227–243.

Foster, G., Holmes, B., Steigerwalt, A. G., Lawson, P. A., Thorne, P., Byrer, D. E., Ross, H. M., Xerry, J., Thompson, P. M. and Collins, M. D. (2004). Campylobacter insulaenigrae sp. nov., isolated from marine mammals, *International Journal of Systematic and Evolutionary Microbiology* **54**(6): 2369–2373.

Fouts, D. E., Mongodin, E. F., Mandrell, R. E., Miller, W. G., Rasko, D. a., Ravel, J., Brinkac, L. M., DeBoy, R. T., Parker, C. T., Daugherty, S. C., Dodson, R. J., Durkin, a. S., Madupu, R., Sullivan, S. a., Shetty, J. U., Ayodeji, M. a., Shvartsbeyn, A., Schatz, M. C., Badger, J. H., Fraser, C. M. and Nelson, K. E. (2005). Major structural differences and novel potential virulence mechanisms from the genomes of multiple campylobacter species, *PLoS Biology* **3**(1): e15.

FSA (2016). Campylobacter contamination in fresh whole chilled UK-produced chickens at retail: January-March 2016, *Technical report*, FSA.

Gallie, J., Libby, E., Bertels, F., Remigi, P., Jendresen, C. B., Ferguson, G. C., Desprat, N., Buffing, M. F., Sauer, U., Beaumont, H. J. E., Martinussen, J., Kilstrup, M. and Rainey, P. B. (2015). Bistability in a Metabolic Network Underpins the De Novo Evolution of Colony Switching in Pseudomonas fluorescens, *PLoS Biology* **13**(3): 1–28.

García-Rodríguez, J. Á. and Fresnadillo Martínez, M. J. (2002). Dynamics of nasopharyngeal colonization by potential respiratory pathogens, *Journal of Antimicrobial Chemotherapy* **50 Suppl.**: 59–73.

Gaynor, E. C., Cawthraw, S., Manning, G., MacKichan, J. K., Falkow, S. and Newell, D. G. (2004). The genome-sequenced variant of Campylobacter jejuni NCTC 11168 and the original clonal clinical isolate differ markedly in colonization, gene expression, and virulence-associated phenotypes, *Journal of Bacteriology* **186**(2): 503–17.

Gebhart, C. J., Edmonds, P., Ward, G. E., Kurtz, H. J. and Brenner, D. O. N. J. (1985). Campylobacter Found in the Intestines of Pigs and Other Animals, **21**(5): 715–720.

Gerlini, A., Colomba, L., Furi, L., Braccini, T., Manso, A. S., Pammolli, A., Wang, B., Vivi, A., Tassini, M., van Rooijen, N., Pozzi, G., Ricci, S., Andrew, P. W., Koedel, U., Moxon, E. R. and Oggioni, M. R. (2014). The Role of Host and Microbial Factors in the Pathogenesis of Pneumococcal Bacteraemia Arising from a Single Bacterial Cell Bottleneck, *PLoS Pathogens* **10**(3): e1004026.

Gilbert, M. J., Kik, M., Miller, W. G., Duim, B. and Wagenaar, J. A. (2015). Campylobacter iguaniorum sp. nov., isolated from reptiles, *International Journal of Systematic and Evolutionary Microbiology* **65**(3): 975–982.

Gillespie, J. H. (1981). Mutation Modification in a Random Environment, *Evolution* **35**(3): 468–476.

González, M., Villanueva, M. P., Debruyne, L., Vandamme, P. and Fernández, H. (2011). Campylobacter insulaenigrae: first isolation report from South American Sea Lion (Otaria flavescens, (Shaw, 1800), *Brazilian Journal of Microbiology* **42**: 261–265.

Grajewski, B. A., Kusek, J. W. and Gelfand, H. M. (1985). Development of a Bacteriophage Typing System for Campylobacter jejuni and Campylobacter coli, *Journal of Clinical Microbiology* **22**(1): 13–18.

Guerry, P., Ewing, C. P., Schirm, M., Lorenzo, M., Kelly, J., Pattarini, D., Majam, G., Thibault, P. and Logan, S. (2006). Changes in flagellin glycosylation affect Campylobacter autoagglutination and virulence, *Molecular Microbiology* **60**(2): 299–311.

Guerry, P. and Szymanski, C. M. (2002). Phase variation of Campylobacter jejuni 81-176 lipooligosaccharide affects ganglioside mimicry and invasiveness in vitro, *Infection and Immunity* **70**(2): 787–793.

Guerry, P. and Szymanski, C. M. (2008). Campylobacter sugars sticking out, *Trends in Microbiology* **16**(9): 428–35.

Gundogdu, O., Bentley, S. D., Holden, M. T., Parkhill, J., Dorrell, N. and Wren, B. W. (2007). Re-annotation and re-analysis of the Campylobacter jejuni NCTC11168 genome sequence, *BMC Genomics* **8**(1): 162.

Gutiérrez-Martín, C. B., Yubero, S., Martínez, S., Frandoloso, R. and Rodríguez-Ferri, E. F. (2011). Evaluation of efficacy of several disinfectants against Campylobacter jejuni strains by a suspension test, *Research in Veterinary Science* **91**(3).

Haber, J. E. (2012). Mating-type genes and MAT switching in Saccharomyces cerevisiae, *Genetics* **191**(1): 33–64.

Handel, A. and Bennett, M. R. (2008). Surviving the bottleneck: transmission mutants and the evolution of microbial populations, *Genetics* **180**(4): 2193–200.

Hansen, V. M., Rosenquist, H., Baggesen, D. L., Brown, S. and Christensen, B. B. (2007). Characterization of Campylobacter phages including analysis of host range by selected Campylobacter Penner serotypes, *BMC Microbiology* **7**: 90.

Hazeleger, W. C., Wouters, J. A., Rombouts, F. M. and Abee, T. (1998). Physiological activity of Campylobacter jejuni far below the minimal growth temperature, *Applied and Environmental Microbiology* **64**(10): 3917–22.

Henderson, I. R., Owen, P. and Nataro, J. P. (1999). Molecular switches–the ON and OFF of bacterial phase variation, *Molecular Microbiology* **33**(5): 919–932.

Hermans, D., Van Deun, K., Martel, A., Van Immerseel, F., Messens, W., Heyndrickx, M., Haesebrouck, F. and Pasmans, F. (2011). Colonization factors of Campylobacter jejuni in the chicken gut, *Veterinary Research* **42**(1): 82.

Hitchen, P., Brzostek, J., Panico, M., Butler, J. A., Morris, H. R., Dell, A. and Linton, D. (2010). Modification of the Campylobacter jejuni flagellin glycan by the product of the Cj1295 homopolymeric-tract-containing gene, *Microbiology* (156): 1953–1962.

Howard, S. L., Jagannathan, A., Soo, E. C., Hui, J. P. M., Aubry, A. J., Ahmed, I., Karlyshev, A., Kelly, J. F., Jones, M. A., Stevens, M. P., Logan, S. M. and Wren, B. W. (2009). Campylobacter jejuni glycosylation island important in cell charge, legionaminic acid biosynthesis, and colonization of chickens, *Infection and Immunity* **77**(6): 2544–2556.

Hughes, R. A. C., Swan, A. V., Raphaël, J.-C., Annane, D., van Koningsveld, R. and van Doorn, P. A. (2007). Immunotherapy for Guillain-Barré syndrome: a systematic review, *Brain: a Journal of Neurology* **130**(Pt 9): 2245–57.

Humphrey, S., Chaloner, G., Kemmett, K., Davidson, N., Williams, N., Kipar, A. and Humphrey, T. (2014). Campylobacter jejuni Is Not Merely a Commensal in Commercial Broiler Chickens and Affects Bird Welfare, *mBio* **5**(4): e01364–14.

Huson, D. H. and Xie, C. (2014). A poor man's blastx-High-throughput metagenomic protein database search using pauda, *Bioinformatics* **30**(1): 38–39.

Jacobs-Reitsma, W., Lyhs, U. and Wagenaar, J. A. (2008). Campylobacter in the Food Supply, *in* I. Nachamkin, C. M. Syzmanski and M. J. Blaser (eds), *Campylobacter, 3rd edition*, ASM Press, pp. 627–644.

Javed, M. A., Grant, A. J., Bagnall, M. C., Maskell, D. J., Newell, D. G. and Manning, G. (2010). Transposon mutagenesis in a hyper-invasive clinical isolate of Campylobacter jejuni reveals a number of genes with potential roles in invasion, *Microbiology* **156**(Pt 4): 1134–43.

Jenny, B. and Kelso, N. V. (2007). Color design for the color vision impaired, *Cartographic Perspectives* **57**: 61–67.

Jerome, J. P., Bell, J. A., Plovanich-Jones, A. E., Barrick, J. E., Brown, C. T. and Mansfield, L. S. (2011). Standing genetic variation in contingency loci drives the rapid adaptation of Campylobacter jejuni to a novel host, *PLoS ONE* **6**(1): e16399.

Jolley, K. A. and Maiden, M. C. J. (2010). BIGSdb: Scalable analysis of bacterial genome variation at the population level, *BMC Bioinformatics* **11**(1): 595.

Kaakoush, N. O. and Mitchell, H. M. (2012). Campylobacter concisus – A New Player in Intestinal Disease, *Frontiers in Cellular and Infection Microbiology* **2**(February): 1–15.

Karlyshev, A. V., Ketley, J. M. and Wren, B. W. (2005). The Campylobacter jejuni glycome, *FEMS Microbiology Reviews* **29**(2): 377–90.

Karlyshev, A. V., Linton, D., Gregson, N. A., Lastovica, A. J. and Wren, B. W. (2000). Genetic and biochemical evidence of a Campylobacter jejuni capsular polysaccharide that accounts for Penner serotype specificity, *Molecular Microbiology* **35**(3): 529–541.

Karlyshev, A. V., Linton, D., Gregson, N. a. and Wren, B. W. (2002). A novel paralogous gene family involved in phase-variable flagella-mediated motility in Campylobacter jejuni., *Microbiology* **148**(Pt 2): 473–80.

Karlyshev, A. V., Mccrossan, M. V. and Brendan, W. (2001). Demonstration of Polysaccharide Capsule in Campylobacter jejuni Using Electron Microscopy, *Infection and Immunity* **69**(9): 5921–5924.

Karlyshev, A. V., Wren, B. W. and Moran, A. P. (2008). Campylobacter jejuni Capsular Polysaccharide, *in* I. Nachamkin, C. M. Szymanski and M. J. Blaser (eds), *Campylobacter, 3rd edition*, 3rd edn, ASM Press, Washington, DC, chapter 28, pp. 505–521.

Kauffman, F. and Mitsui, C. (1930). Zwei neue Paratyphustypen mit bisher unbekanntem Phasenwechsel, *Zeitschrift für Hygiene und Infektionskrankheiten* **111**(5): 640–648.

Kielbasa, S. M., Wan, R., Sato, K., Kiebasa, S. M., Horton, P. and Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison, *Genome Research* pp. 487–493.

Kienesberger, S., Sprenger, H., Wolfgruber, S., Halwachs, B., Thallinger, G. G., Perez-Perez, G. I., Blaser, M. J., Zechner, E. L. and Gorkiewicz, G. (2014). Comparative genome analysis of Campylobacter fetus subspecies revealed horizontally acquired genetic elements important for virulence and niche specificity, *PLoS ONE* **9**(1).

Kim, J.-S., Artymovich, K. A., Hall, D. F., Smith, E. J., Fulton, R., Bell, J., Dybas, L., Mansfield, L. S., Tempelman, R., Wilson, D. L. and Linz, J. E. (2012). Passage of Campylobacter jejuni through the chicken reservoir or mice promotes phase variation in contingency genes Cj0045 and Cj0170 that strongly associates with colonization and disease in a mouse model, *Microbiology* **158**(Pt 5): 1304–16.

King, E. O. (1957). Human Infections with Vibrio fetus and a closely related vibrio, *The Journal of Infectious Diseases* **101**(2): 119–128.

Klar, A. J. S. (2010). The yeast mating-type switching mechanism: A memoir, *Genetics* **186**(2): 443–449.

Kozbial, P. Z. and Mushegian, A. R. (2005). Natural history of S-adenosylmethionine-binding proteins., *BMC Structural Biology* **5**: 19.

Kulldorff, M., Fang, Z. and Walsh, S. J. (2003). A Tree-Based Scan Statistic for Database Disease Surveillance, *Biometrics* **59**(2): 323–331.

Lango-Scholey, L., Aidley, J., Woodacre, A., Jones, M. A. and Bayliss, C. D. (2016). High Throughput Method for Analysis of Repeat Number for 28 Phase Variable Loci of Campylobacter jejuni Strain NCTC 11168, *PLOS ONE* **11**(7): e0159634.

Lawson, A. J., Linton, D. and Stanley, J. (1998). 16S rRNA gene sequences of 'Candidatus Campylobacter hominis', a novel uncultivated species, are found in the gastrointestinal tract of healthy humans, *Microbiology* **144**(8): 2063–2071.

Lederberg, J. and Iino, T. (1956). Phase Variation in Salmonella, *Genetics* **41**(5): 743–757.

Lee, J. W. and Helmann, J. D. (2007). Functional specialization within the fur family of metalloregulators, *BioMetals* **20**(3-4): 485–499.

Leigh, E. G. (1970). Natural Selection and Mutability, *The American Naturalist* **104**(937): 301–305.

Levinson, G. and Gutman, G. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution, *Molecular Biology and Evolution* **4**(3): 203–221.

Libby, E. and Rainey, P. B. (2011). Exclusion rules, bottlenecks and the evolution of stochastic phenotype switching, *Proceedings. Biological Sciences / The Royal Society* **278**(1724): 3574–83.

Lin, W.-H. and Kussell, E. (2012). Evolutionary pressures on simple sequence repeats in prokaryotic coding regions, *Nucleic Acids Research* **40**(6): 2399–413.

Linton, D., Gilbert, M., Hitchen, P. G., Dell, A., Morris, H. R., Wakarchuk, W. W., Gregson, N. a. and Wren, B. W. (2000). Phase variation of a beta-1,3 galactosyltransferase involved in generation of the ganglioside GM1-like lipo-oligosaccharide of Campylobacter jejuni, *Molecular Microbiology* **37**(3): 501–14.

Loessner, I., Dietrich, K., Dittrich, D., Hacker, J. and Ziebuhr, W. (2002). Transposase-dependent formation of circular IS256 derivatives in Staphylococcus epidermidis and Staphylococcus aureus, *Journal of Bacteriology* **184**(17): 4709–4714.

Lotka, A. J. (1910). Contribution to the theory of periodic reactions, *Journal of Physcial Chemistry* **14**(3): 271–274.

Lotka, A. J. (1920). Analytical Note on Certain Rhythmic Relations in Organic Systems, *Proceedings of the National Academy of Sciences of the United States of America* **6**(7): 410–415.

Lynett, J. (1999). *Defining the Role of CipA in the Pathogenesis of Campylobacter jejuni Infection*, Msc thesis, University of Toronto.

Macé, S., Haddad, N., Zagorec, M. and Tresse, O. (2015). Influence of measurement and control of microaerobic gaseous atmospheres in methods for Campylobacter growth studies, *Food Microbiology* **52**: 169–176.

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., Feavers, I. M., Achtman, M. and Spratt, B. G. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms, *Proceedings of the National Academy of Sciences of the United States of America* **95**(6): 3140–5.

Maiden, M. C. J., Jansen van Rensburg, M. J., Bray, J. E., Earle, S. G., Ford, S. a., Jolley, K. a. and McCarthy, N. D. (2013). MLST revisited: the gene-by-gene approach to bacterial genomics, *Nature Reviews Microbiology* **11**(10): 728–36.

Manso, A. S., Chai, M. H., Atack, J. M., Furi, L., De Ste Croix, M., Haigh, R., Trappetti, C., Ogunniyi, A. D., Shewell, L. K., Boitano, M., Clark, T. a., Korlach, J., Blades, M., Mirkes, E., Gorban, A. N., Paton, J. C., Jennings, M. P. and Oggioni, M. R. (2014). A random six-phase switch regulates pneumococcal virulence via global epigenetic changes, *Nature Communications* **5**: 5055.

Martinot, M., Jaulhac, B., Moog, R., De Martino, S., Kehrli, P., Monteil, H. and Piemont, Y. (2001). Campylobacter lari bacteremia, *Clinical Microbiology and Infection* **7**(2): 96–97.

Maue, A. C., Mohawk, K. L., Giles, D. K., Poly, F., Ewing, C. P., Jiao, Y., Lee, G., Ma, Z., Monteiro, M. A., Hill, C. L., Ferderber, J. S., Porter, C. K. and Trent, M. S. (2013). The Polysaccharide Capsule of Campylobacter jejuni Modulates the Host Immune Response, *Infection and Immunity* **81**(3): 665–672.

McDonald, S. and Mautner, L. S. (1970). A case of human vibriosis, *Canadian Medical Association Journal* **103**(9): 951–952.

McNally, D. J., Lamoureux, M. P., Karlyshev, A. V., Fiori, L. M., Li, J., Thacker, G., Coleman, R. A., Khieu, N. H., Wren, B. W., Brisson, J.-R., Jarrell, H. C. and Szymanski, C. M. (2007). Commonality

and biosynthesis of the O-methyl phosphoramidate capsule modification in Campylobacter jejuni, *The Journal of Biological Chemistry* **282**(39): 28566–76.

Merkel, A. and Gemmell, N. (2008). Detecting short tandem repeats from genome data: opening the software black box, *Briefings in Bioinformatics* **9**(5): 355–366.

Merriam, C. V., Fernandez, H. T., Citron, D. M., Tyrrell, K. L., Warren, Y. A. and Goldstein, E. J. C. (2003). Bacteriology of human bite wound infections, *Anaerobe* **9**(2): 83–86.

Meynell, G. G. and Maw, J. (1968). Evidence for a two-stage model of microbial infection, *The Journal of Hygiene* **66**(2): 273–280.

Middelkamp, N. J. and Wolf, H. A. (1961). Infection due to a "related" Vibrio, *The Journal of Pediatrics* **59**(3): 318–321.

Miller, W. G., Yee, E., Chapman, M. H., Smith, T. P. L., Bono, J. L., Huynh, S., Parker, C. T., Vandamme, P., Luong, K. and Korlach, J. (2014). Comparative genomics of the Campylobacter lari group, *Genome Biology and Evolution* **6**(12): 3252–3266.

Mitrophanov, A. Y. A. and Groisman, E. E. a. (2008). Signal integration in bacterial two-component regulatory systems., *Genes & development* **22**(19): 2601–2611.

Mittenhuber, G. (2002). An inventory of genes encoding RNA polymerase sigma factors in 31 completely sequenced eubacterial genomes, *Journal of Molecular Microbiology and Biotechnology* **4**(1): 77–91.

Moran, A. P. and Penner, J. L. (1999). Serotyping of Campylobacter jejuni based on heat-stable antigens: Relevance, molecular basis and implications in pathogenesis, *Journal of Applied Microbiology* **86**(3): 361–377.

Morel, P., Reverdy, C., Michel, B., Ehrlich, S. D. and Cassuto, E. (1998). The role of SOS and flap processing in microsatellite instability in Escherichia coli, *Proceedings of the National Academy of Sciences of the United States of America* **95**(17): 10003–10008.

Moxon, R., Bayliss, C. D. and Hood, D. (2006). Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation, *Annual Review of Genetics* **40**: 307–33.

Moxon, R. and Murphy, P. (1978). Haemophilus influenzae bacteremia and meningitis resulting from survival of a single organism, *Proceedings of the National Academy of Sciences of the United States of America* **75**(3): 1534–1536.

Mrázek, J., Guo, X. and Shah, A. (2007). Simple sequence repeats in prokaryotic genomes, *Proceedings of the National Academy of Sciences of the United States of America* **104**(20): 8472–8477.

Murray, I. A., Clark, T. A., Morgan, R. D., Boitano, M., Anton, B. P., Luong, K., Fomenkov, A., Turner, S. W., Korlach, J. and Roberts, R. J. (2012). The methylomes of six bacteria, *Nucleic Acids Research* **40**(22): 11450–62.

Murrell, P. (2005). *R Graphics*, 1st edn, CRC Press, Boca Raton, USA.

Nachamkin, I., Allos, B. M. and Ho, T. (1998). Campylobacter species and Guillain-Barré syndrome, *Clinical Microbiology Reviews* **11**(3): 555–67.

Nachamkin, I., Liu, J., Li, M., Ung, H., Moran, A. P., Prendergast, M. M. and Sheikh, K. (2002). Campylobacter jejuni from Patients with Guillain-Barré Syndrome Preferentially Expresses a GD1a-Like Epitope, *Infection and Immunity* **70**(9): 5299–5303.

Nakase, Y., Takatsu, K. and Kasuga, T. (1969). Antigenic Structure and Phase Bordetella pertussis Variation in, *Japanese Journal of Microbiology* **13**(3): 283–291.

O'Donovan, D., Corcoran, G. D., Lucey, B. and Sleator, R. D. (2014). Campylobacter ureolyticus: a portrait of the pathogen, *Virulence* **5**(4): 498–506.

O'Leary, J., Corcoran, D. and Lucey, B. (2009). Comparison of the EntericBio multiplex PCR system with routine culture for detection of bacterial enteric pathogens, *Journal of Clinical Microbiology* **47**(11): 3449–3453.

Orsi, R. H., Bowen, B. M. and Wiedmann, M. (2010). Homopolymeric tracts represent a general regulatory mechanism in prokaryotes, *BMC Genomics* **11**: 102.

Palmer, M. E., Lipsitch, M., Moxon, E. R. and Bayliss, C. D. (2013). Broad Conditions Favor the Evolution of Phase-Variable Loci, *mBio* **4**(1): 1–9.

Parkhill, J., Wren, B. W., Mungall, K., Ketley, J. M., Churcher, C., Basham, D., Chillingworth, T., Davies, R. M., Feltwell, T., Holroyd, S., Jagels, K., Karlyshev, a. V., Moule, S., Pallen, M. J., Penn, C. W., Quail, M. a., Rajandream, M. a., Rutherford, K. M., van Vliet, a. H., Whitehead, S. and Barrell, B. G. (2000). The genome sequence of the food-borne pathogen Campylobacter jejuni reveals hypervariable sequences, *Nature* **403**(6770): 665–8.

Pavelka, M. S., Wright, L. F. and Silver, R. P. (1991). Identification of two genes, kpsM and kpsT, in region 3 of the polysialic acid gene cluster of Escherichia coli K1, *Journal of Bacteriology* **173**(15): 4603–4610.

Pearson, B. M., Gaskin, D. J. H., Segers, R. P. a. M., Wells, J. M., Nuijten, P. J. M. and van Vliet, A. H. M. (2007). The complete genome sequence of Campylobacter jejuni strain 81116 (NCTC 11828), *Journal of Bacteriology* **189**(22): 8402–3.

Pearson, B. M., Rokney, A., Crossman, L. C., Miller, W. G., Wain, J. and van Vliet, A. H. M. (2013). Complete Genome Sequence of the Campylobacter coli Clinical Isolate 15-537360, *Genome Announcements* **1**(6): e01056–13.

Pearson, W. R. (2013). Selecting the Right Similarity-Scoring Matrix, *Current Protocols in Bioinformatics* **43**: 3.5.1–3.5.9.

Pérez-Losada, M., Cabezas, P., Castro-Nallar, E. and Crandall, K. A. (2013). Pathogen typing in the genomics era: MLST and the future of molecular epidemiology, *Infection, Genetics and Evolution* **16**: 38–53.

Perkins-Balding, D., Duval-Valentin, G. and Glasgow, A. C. (1999). Excision of IS492 requires flanking target sequences and results in circle formation in Pseudoalteromonas atlantica, *Journal of Bacteriology* **181**(16): 4937–4948.

Petersen, L., Larsen, T. S., Ussery, D. W., On, S. L. W. and Krogh, A. (2003). Rpo D promoters in Campylobacter jejuni exhibit a strong periodic signal instead of a - 35 box, *Journal of Molecular Biology* **326**(5): 1361–1372.

R Core Team (2016). R: A Language and Environment for Statistical Computing.
URL: *http://www.r-project.org/*

Rajewsky, K. (1996). Clonal selection and learning in the antibody system, *Nature* **381**(6585): 751–758.

Revez, J., Schott, T., Rossi, M. and Hänninen, M. L. (2012). Complete genome sequence of a variant of campylobacter jejuni NCTC 11168, *Journal of Bacteriology* **194**(22): 6298–6299.

Ritchie, A. E., Bryner, J. H. and Foley, J. W. (2016). Role of DNA and bacteriophage in campylobacter auto-agglutination, *Journal of Medical Microbiology* **8**(1983): 333–340.

Rost, B. (2002). Enzyme function less conserved than anticipated, *Journal of Molecular Biology* **318**(2): 595–608.

Rubin, L. G. (1987). Bacterial Colonization and Infection Resulting from Multiplication of a Single Organism, *Clinical Infectious Diseases* **9**(3): 488–493.

Ryan, K. A. and Lo, R. Y. C. (1999). Characterization of a CACAG pentanucleotide repeat in Pasteurella haemolytica and its possible role in modulation of a novel type III restriction-modification system, *Nucleic Acids Research* **27**(6): 1505–1511.

Salathé, M., Van Cleve, J. and Feldman, M. W. (2009). Evolution of stochastic switching rates in asymmetric fitness landscapes, *Genetics* **182**(4): 1159–64.

Saunders, N. J., Peden, J. F., Hood, D. W. and Moxon, E. R. (1998). Simple sequence repeats in the Helicobacter pylori genome, *Molecular Microbiology* **27**(6): 1091–1098.

Scherer, R. (2014). PropCIs: Various confidence interval methods for proportions.
URL: *http://cran.r-project.org/package=PropCIs*

Schnoes, A. M., Brown, S. D., Dodevski, I. and Babbitt, P. C. (2009). Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies, *PLoS Computational Biology* **5**(12).

Sebald, M. and Véron, M. (1963). Teneur en bases de 1'ADN et classification des vibrions, *Annales de l'Institut Pasteur* **105**: 897–910.

Seib, K. L., Peak, I. R. A. and Jennings, M. P. (2002). Phase variable restriction-modification systems in Moraxella catarrhalis, *FEMS Immunology and Medical Microbiology* **32**: 159–165.

Shannon, C. E. (1948). A Mathematical Theory of Communication, *Bell System Technical Journal* **27**(3): 379–423.

Shyaka, A., Kusumoto, A., Asakura, H. and Kawamoto, K. (2015). Whole-Genome Sequences of Eight Campylobacter jejuni Isolates from, *Genome Announcements* **3**(2): e00315–15.

Silipo, A., Molinaro, A., Sturiale, L., Dow, J. M., Erbs, G., Lanzetta, R., Newman, M. A. and Parrilli, M. (2005). The elicitation of plant innate immunity by lipooligosaccharide of Xanthomonas campestris, *Journal of Biological Chemistry* **280**(39): 33660–33668.

Skarp, C. P. A., Akinrinade, O., Nilsson, A. J. E., Ellström, P., Myllykangas, S. and Rautelin, H. (2015). Comparative genomics and genome biology of invasive Campylobacter jejuni, *Scientific Reports* **5**: 17300.

Skirrow, M. B. (1977). Campylobacter enteritis: a "new" disease, *British Medical Journal* (2): 9–11.

Skirrow, M. B. (2006). John McFadyean and the centenary of the first isolation of Campylobacter species, *Clinical Infectious Diseases* **43**(9): 1213–7.

Smith, T. and Taylor, M. S. (1919). Some morphological and biological characters of the spirilla (Vibrio fetus, N. Sp.) associated with disease of the fetal membrane in cattle, *The Journal of Experimental Medicine* **30**(4): 299–311.

Snelling, W. J., Matsuda, M., Moore, J. E. and Dooley, J. S. G. (2005). Campylobacter jejuni, *Letters in Applied Microbiology* **41**(4): 297–302.

Sørensen, M. C. H., Gencay, Y. E., Birk, T., Baldvinsson, S. B., Jäckel, C., Hammerl, J. a., Vegge, C. S., Neve, H. and Brøndsted, L. (2015). Primary Isolation Strain Determines Both Phage Type and Receptors Recognised by Campylobacter jejuni Bacteriophages, *PloS ONE* **10**(1): e0116287.

Sørensen, M. C. H., van Alphen, L. B., Fodor, C., Crowley, S. M., Christensen, B. B., Szymanski, C. M. and Brøndsted, L. (2012). Phase variable expression of capsular polysaccharide modifications allows Campylobacter jejuni to avoid bacteriophage infection in chickens, *Frontiers in Cellular and Infection Microbiology* **2**(February): 11.

Sørensen, M. C. H., van Alphen, L. B., Harboe, A., Li, J., Christensen, B. B., Szymanski, C. M. and Brøndsted, L. (2011). Bacteriophage F336 recognizes the capsular phosphoramidate modification of Campylobacter jejuni NCTC11168, *Journal of Bacteriology* **193**(23): 6742–9.

Srikhanta, Y. N., Fox, K. L. and Jennings, M. P. (2010). The phasevarion: phase variation of type III DNA methyltransferases controls coordinated switching in multiple genes, *Nature reviews. Microbiology* **8**(3): 196–206.

Srikhanta, Y. N., Gorrell, R. J., Steen, J. A., Gawthorne, J. A., Kwok, T., Grimmond, S. M., Robins-Browne, R. M. and Jennings, M. P. (2011). Phasevarion mediated epigenetic gene regulation in Helicobacter pylori, *PloS ONE* **6**(12): e27569.

Srikhanta, Y. N., Maguire, T. L., Stacey, K. J., Grimmond, S. M. and Jennings, M. P. (2005). The phasevarion: a genetic system controlling coordinated, random switching of expression of multiple genes, *Proceedings of the National Academy of Sciences of the United States of America* **102**(15): 5547–51.

Sternberg, M. J. E., Tamaddoni-Nezhad, A., Lesk, V. I., Kay, E., Hitchen, P. G., Cootes, A., Van Alphen, L. B., Lamoureux, M. P., Jarrell, H. C., Rawlings, C. J., Soo, E. C., Szymanski, C. M., Dell, A., Wren, B. W. and Muggleton, S. H. (2013). Gene Function Hypotheses for the Campylobacter jejuni Glycome Generated by a Logic-Based Approach, *Journal of Molecular Biology* **425**(1): 186–197.

Stocker, B. A. (1949). Measurements of rate of mutation of flagellar antigenic phase in Salmonella typhimurium, *The Journal of Hygiene* **47**(4): 398–413.

Stoddard, R. A., Miller, W. G., Foley, J. E., Lawrence, J., Gulland, F. M. D., Conrad, P. A. and Byrne, B. A. (2007). Campylobacter insulaenigrae isolates from northern elephant seals (Mirounga angustirostris) in California, *Applied and Environmental Microbiology* **73**(6): 1729–1735.

Stoker, M. G. P. and Fiset, P. (1956). Phase variation in the Nine Mile and other strains of Rickettsia burneti, *Canadian Journal of Microbiology* **2**(3): 310–321.

Svensson, S. L., Hyunh, S., Parker, C. T. and Gaynor, E. C. (2015). The Campylobacter jejuni CprRS two-component regulatory system regulates aspects of the cell envelope, *Molecular Microbiology* **96**(1): 189–209.

Tanner, A. C. R., Badger, S., Lai, C., Listgarten, M. A. X. A., Visconti, R. A. and Socransky, S. S. (1981). Wollinella gen. nov. succinogenes (Vibrio succinogenes Wolin et al.) comb. nov., and Description of Bacteroides gracilis sp. nov., Wolinela recta sp. nov., Campylobacter concisus sp. nov., and Eikenella corrodens from Humans with Periodontal Disease, *International Journal of Systematic Bacteriology* **31**(4): 432–445.

Tanner, A. C. R., Listgarten, M. A. and Ebersole, J. L. (1984). Wolinella curva sp. nov.: "Vibrio succinogenes" of Human Origin, *International Journal of Systematic Bacteriology* **34**(3): 275–282.

Tantau, T. (2013). *Graph Drawing in TikZ*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 517–528.

Thomas, D. K., Lone, A. G., Brent Selinger, L., Taboada, E. N., Uwiera, R. R. E., Wade Abbott, D. and Douglas Inglis, G. (2014). Comparative variation within the genome of Campylobacter jejuni NCTC 11168 in human and murine hosts, *PLoS ONE* **9**(2): 1–11.

Thomas, M. T., Shepherd, M., Poole, R. K., Van Vliet, A. H. M., Kelly, D. J. and Pearson, B. M. (2011). Two respiratory enzyme systems in Campylobacter jejuni NCTC 11168 contribute to growth on l-lactate, *Environmental Microbiology* **13**(1): 48–61.

Tonegawa, S. (1983). Somatic generation of antibody diversity, *Nature* **302**(5909): 575–581.

Topley, W. W. C. and Ayrton, J. (1924). Further Investigations into the Biological Characteristics of B. enteritidis (aertrycke), *Journal of Hygiene* **23**(2): 198–222.

Ulasi, G. N., Creese, A. J., Hui, S. X., Penn, C. W. and Cooper, H. J. (2015). Comprehensive mapping of O-glycosylation in flagellin from Campylobacter jejuni 11168: A multienzyme differential ion mobility mass spectrometry approach, *Proteomics* **15**(16): 2733–2745.

van Alphen, L. B., Wenzel, C. Q., Richards, M. R., Fodor, C., Ashmus, R. a., Stahl, M., Karlyshev, A. V., Wren, B. W., Stintzi, A., Miller, W. G., Lowary, T. L. and Szymanski, C. M. (2014). Biological roles of the O-methyl phosphoramidate capsule modification in Campylobacter jejuni, *PLOS ONE* **9**(1): e87051.

van Alphen, L. B., Wuhrer, M., Bleumink-Pluym, N. M. C., Hensbergen, P. J., Deelder, A. M. and van Putten, J. P. M. (2008). A functional Campylobacter jejuni maf4 gene results in novel glycoforms on flagellin and altered autoagglutination behaviour, *Microbiology* **154**(Pt 11): 3385–97.

Van Der Walt, S., Colbert, S. C. and Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation, *Computing in Science and Engineering* **13**(2): 22–30.

van der Woude, M. W. and Bäumler, A. J. (2004). Phase and antigenic variation in bacteria, *Clinical Microbiology Reviews* **17**(3): 581–611.

van Ham, S. M., van Alphen, L., Mooi, F. R. and van Putten, J. P. M. (1993). Phase variation of H. influenzae fimbriae: Transcriptional control of two divergent genes through a variable combined promoter region, *Cell* **73**(6): 1187–1196.

van Vliet, a. H. and Ketley, J. M. (2001). Pathogenesis of enteric Campylobacter infection., *Journal of Applied Microbiology* **90**: 45S–56S.

van Vliet, A. H. M., Ketley, J. M., Park, S. F. and Penn, C. W. (2002). The role of iron in Campylobacter gene regulation, metabolism and oxidative stress defense, *FEMS Microbiology Reviews* **26**(2): 173–186.

van Vliet, A. H. M., Wooldridge, K. G. and Ketley, J. M. (1998). Iron-Responsive Gene Regulation in a Campylobacter jejuni fur Mutant, *Journal of Bacteriology* **180**(20): 5291–5298.

Veazie, L. (1949). A New Type of Phase Variation in the 103 Race of Shigella paradysenteriae, *The Journal of Immunology* **61**: 307–314.

Véron, M. and Chatelain, R. (1973). Taxonomic Study of the Genus Campylobacter Sebald and Vkron and Designation of the Neotype Strain for the Type Species, Campylobacter fetus (Smith and Taylor) Sebald and Véron, *International Journal of Systematic Bacteriology* **23**(2): 122–134.

Vinogradov, E., Kubler-Kielb, J. and Korenevsky, A. (2008). The structure of the carbohydrate backbone of the LPS from Shewanella spp. MR-4, *Carbohydrate Research* **343**(15): 2701–2705.

Wahl, L. M., Gerrish, P. J. and Saika-Voivod, I. (2002). Evaluating the impact of population bottlenecks in experimental evolution, *Genetics* **162**(2): 961–71.

Ward, B. Q. (1948). The Apparent Involvement of Vibrio fetus in an Infection of Man, *Journal of Bacteriology* **55**(1): 113–114.

Wawrzynow, A., Wojtkowiak, D., Marszalek, J., Banecki, B., Jonsen, M., Graves, B., Georgopoulos, C. and Zylicz, M. (1995). The ClpX heat-shock protein of Escherichia coli, the ATP-dependent

substrate specificity component of the ClpP-ClpX protease, is a novel molecular chaperone, *The EMBO Journal* **14**(9): 1867–1877.

Wearne, S. (2013). A refreshed strategy to reduce Campylobacteriosis from poultry, *Technical report*, FSA.

Werno, A. M., Klena, J. D., Shaw, G. M., Murdoch, R. and Murdoch, D. R. (2002). Fatal Case of Campylobacter lari Prosthetic Joint Infection and Bacteremia in an Immunocompetent Patient Fatal Case of Campylobacter lari Prosthetic Joint Infection and Bacteremia in an Immunocompetent Patient, *Journal of Clinical Microbiology* **40**(3): 40–43.

Whisstock, J. C. and Lesk, A. M. (2003). Prediction of protein function from protein sequence and structure, *Quarterly Reviews of Biophysics* **36**(3): 307–340.

Whitfield, C. (2006). Biosynthesis and Assembly of Capsular Polysaccharides, *Annual Review of Biochemistry* **75**: 39–68.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*, Springer New York, New York, USA.

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data, *Journal of Statistical Software* **40**(1): 1–29.

Wilson, C. A., Kreychman, J. and Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores, *Journal of Molecular Biology* **297**(1): 233–49.

Wooldridge, K. G., Williams, P. H. and Ketley, J. M. (1994). Iron-Responsive Genetic, *Journal of Bacteriology* **176**(18): 5852–5856.

Wösten, M. M. S. M. (1998). Eubacterial sigma-factors, *FEMS Microbiology Reviews* **22**: 127–150.

Yang, Q.-L. and C., G. E. (1996). Variation of Gonococcal Lipooligosaccharide Structure Is Due to Alterations in Poly-G Tracts in lgt Genes Encoding Glycosyl Transferases, *Journal of Experimental Medicine* **183**: 323–327.

Yanisch-Perron, C., Vieira, J. and Messing, J. (1985). Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mpl8 and pUC19 vectors, *Gene* **33**(1): 103–119.

Yu, R. K., Usuki, S. and Ariga, T. (2006). Ganglioside molecular mimicry and its pathological roles in Guillain-Barré syndrome and related diseases, *Infection and Immunity* **74**(12): 6517–27.

Yuan, L., Chan, G. C., Beeler, D., Janes, L., Spokes, K. C., Dharaneeswaran, H., Mojiri, A., Adams, W. J., Sciuto, T., Garcia-Cardeña, G., Molema, G., Kang, P. M., Jahroudi, N., Marsden, P. A., Dvorak, A., Regan, E. R. and Aird, W. C. (2016). A role of stochastic phenotype switching in generating mosaic endothelial cell heterogeneity, *Nature Communications* **7**: 10160.

Yuki, N. and Hartung, H.-P. (2012). Guillain-Barré syndrome, *The New England Journal of Medicine* **366**(24): 2294–304.

Zhang, M., He, L., Li, Q., Sun, H., Gu, Y., You, Y., Meng, F. and Zhang, J. (2010). Genomic characterization of the Guillain-Barré syndrome-associated Campylobacter jejuni ICDCCJ07001 Isolate, *PloS ONE* **5**(11): e15060.

Zieg, J., Silverman, M., Hilmen, M. and Simon, M. (1977). Recombinational Switch for Gene Expression, *Science* **196**: 170–172.