GENETIC EPIDEMIOLOGY AND HETEROGENEITY OF *CAMPYLOBACTER SPP.*

STEVEN J. DUNN

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS OF NOTTINGHAM TRENT UNIVERSITY FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

JUNE 2017

Dedicated to my family.

For everything,
Thank you.

Abstract

Initially, this work examines clinical *Campylobacter* isolates obtained from a single health trust site in Nottingham. These results reveal novel sequence types, and identified a previously undescribed peak in incidence that is observable across national data. By utilising a read mapping approach in combination with existing comparative methods, the first instance of case linkage between sporadic clinical isolates was demonstrated. This dataset also revealed an instance of repeat patient sampling, with the resulting isolates showing a marked level of diversity. This generated questions as to whether the diversity that *Campylobacter* exhibits can be resolved to an intra-population level.

To study this further, isolates from the dominant clinical lineage – *C. jejuni* ST-21 – were analysed using a deep sequencing methodology. These results reveal a number of minor allele variations in chemotaxis, membrane and flagellar associated loci, which are hypothesised to undergo variation in response to selective pressures in the human gut. In an expansion of this work, additional datasets were generated to include other clinically relevant CC's, including the *C. coli* lineage ST-828. These results revealed a similar pattern of diversity, with common loci identified as undergoing variation in multiple samples, and in some instances the same amino acid residue. In general, ST-21 isolates exhibited more non-synonymous mutations, distributed across fewer loci, suggesting that this lineage may have a repertoire of alleles that is more adapted to the human host.

To investigate whether the observed diversity is generated during infection, or is maintained from a diverse infectious source, isolates from fresh retail chicken were analysed. These samples revealed a marked decrease in overall diversity, as well as differences in the functions of variable loci. These results show that infections occurring from the consumption of retail chicken (the largest single source of campylobacteriosis) arise from genetically uniform populations, and that the diversity observed in clinical cases is generated within - and furthermore may be specific to - the human host.

Table of Contents

Abstract……iii
List of Figures……vii
List of Tables……ix
List of Abbreviations……x

1. Chapter One: Introduction……1

1.1. History……2

1.2. *Campylobacteraceae*……6

1.2.1. Thermotolerant Species……9

1.3. Clinical manifestations of *Campylobacter*……12

1.3.1. Campylobacteriosis……12

1.3.2. Extraintestinal Complications……14

1.3.3. Guillain-Barré Syndrome……15

1.3.4. Transmission……15

1.3.5. Incidence and Prevalence……20

1.3.6. Pathogenesis……27

1.3.7. Genomics……33

1.3.8. Study Rationale……38

2. Chapter Two: Methods……40

2.1. Bacterial Strains……41

2.2. Sampling……41

2.3. Culture Media……43

2.3.1. Culture Conditions……43

2.4. DNA Isolation……46

2.4.1. DNA Qualification……48

2.4.2. DNA Quantification……49

2.5. Whole Genome Sequencing……49

2.5.1. Estimating Sequencing Coverage……51

2.5.2. Genomic Read Curation……52

2.5.3. Genome Assembly and Annotation……52

2.5.4. MLST Profiling……53

2.5.5. Core Genome Phylogeny……53

2.5.6. Deep Sequencing……54

3. Chapter Three: Genetic epidemiology of clinical *Campylobacter spp.* 56
3.1. Abstract 57
3.2. Introduction 59
3.2.1. Comparative genomics of *Campylobacter spp.* 61
3.3. Methods 65
3.4. Results 67
3.4.1. Incidence 67
3.4.2. Sequencing and Assembly Quality 74
3.4.3. MLST 75
3.4.4. Phylogeny 77
3.4.5. Single Patient Isolates 82
3.5. Discussion 83
4. Chapter Four: Intrapopulation diversity amongst clinical *C. jejuni* isolates 94
4.1. Abstract 95
4.2. Introduction 97
4.2.1. Bacterial evolution 97
4.2.2. Within-host variation 98
4.2.3. Limitations of assembly based genomics 100
4.2.4. Measuring intra-population diversity 101
4.2.5. The Diversity of *Campylobacter spp.* 103
4.3. Methods 105
4.3.1. Sequencing 105
4.3.2. Read Mapping 106
4.3.3. Variant Calls 107
4.3.4. Statistical Analysis of Assembly Metrics 108
4.4. Results 109
4.4.1. Isolates, sequencing and assembly. 109
4.4.2. Minor allele variation 111
4.4.3. Chemotaxis 112
4.4.4. Flagellar and Membrane 116
4.4.5. Homopolymeric Tract 118
4.4.6. Intergenic 118
4.4.7. Hypervariable Loci 119
4.4.8. Loci in relation to NCTC11168 120

4.5. Discussion 121
4.5.1. Chemotaxis Genes 122
4.5.2. Membrane and Flagella 127
4.5.3. Homopolymeric Tracts 130
4.5.4. Future work 132
5. Chapter Five: Genetic heterogeneity of *Campylobacter spp.* 133
5.1. Abstract 134
5.2. Introduction 136
5.2.1. Campylobacteriosis and Poultry 137
5.3. Methods 138
5.3.1. Poultry 138
5.3.2. Clinical Isolates 139
5.3.3. Deep Sequencing 139
5.4. Results 140
5.4.1. Non ST-21 Isolates 140
5.4.2. Poultry Isolates 151
5.4.3. Cumulative Statistics 157
5.4.4. Comparison of Diversity 159
5.4.5. All loci relative to NCTC11168 160
5.5. Discussion 163
5.5.1. Non ST-21 Isolates 163
5.5.2. Poultry Dataset 169
5.5.3. Comparing deep sequence datasets 172
5.5.4. Future Work 175
6. Chapter Six: Conclusion 177
7. Acknowledgements 184
8. References 186
9. Appendix 225

List of Figures

Figure 1.1 - Number of publications* with '*Campylobacter'* in the title before and after the introduction of Skirrow's Supplement 5
Figure 1.2 - Phylogenetic analysis of *Campylobacter* 16S rRNA regions 8
Figure 1.3 - Incidence rates of *Campylobacter* infections in European countries based on individually calculated reporting ratios* 22
Figure 1.4 - A brief overview of *Campylobacter* pathogenesis 32
Figure 1.5 - Total number of isolates from identified clonal complexes submitted to BigsDB isolated from human stool 36
Figure 2.1- Summary of sample preparation and DNA qualification process. 45
Figure 2.2 - Qiagen QIAmp DNA Mini Kit DNA Isolation Workflow 47
Figure 2.3 - Overview of the deep sequencing analysis pipeline 55
Figure 3.1 - Monthly incidence of campylobacteriosis cases identified across the Queen's Medical Centre's catchment area during 2014 67
Figure 3.2 - Current vs. historical incidence of campylobacteriosis. within the greater Nottinghamshire area 68
Figure 3.3 - Incidence of campylobacteriosis across Nottinghamshire during February, 2014 69
Figure 3.4 - Percentage of average *Campylobacter spp*. incidence across the United Kingdom, from 2008-2011* 70
Figure 3.5 – Monthly distribution of campylobacteriosis cases amongst age groups in the greater Nottinghamshire area 71
Figure 3.6 - Monthly incidence of campylobacteriosis across the greater Nottinghamshire area with abundance of constituent patient age groups 72
Figure 3.7 - Distribution of campylobacteriosis incidence* across Nottinghamshire postcode areas during 2014 73
Figure 3.8 – Distribution of quality metrics amongst the final genomic dataset, as reported by Quast 74
Figure 3.9 - Distribution of clonal complexes amongst 158 clinical *Campylobacter spp*. isolates obtained from Queen's Medical Centre 76
Figure 3.10 - Core genome phylogeny of 143 clinical *C. jejuni* isolates 78
Figure 3.11 - Core genome phylogeny of isolates belonging to clonal complex ST-21 with linked isolate pairs and respective number of Single Nucleotide Polymorphisms (SNPs) highlighted 79
Figure 3.12 - Core genome phylogeny of clinical *C. coli* isolates 80
Figure 3.13 - The average monthly temperature and rainfall across Nottinghamshire from 2012-2014 84
Figure 4.1 - Curation and mapping pipeline for the ST-21 deep sequencing dataset 107
Figure 4.2 - Quality metrics of draft *de novo* assemblies 110
Figure 4.3 - The statistical impact of sequencing and assembly metrics on the number of loci in which minor allele variation was detected 110
Figure 4.4 - Topology of Tlp8 (CetZ) encoded by the CAMP1031/Cj1110c locus 114
Figure 4.5 - Topology of CetA encoded by the CAMP1109/Cj1190c locus 114
Figure 4.6- Topology of CAMP1460 (CcmL) highlighted with positions of variable residues. 115
Figure 4.7 - Circular chromosome of NCTC111168 with annotated CDS* 120
Figure 5.1 - Quality metrics of draft *de novo* assemblies of non ST-21 isolates 141
Figure 5.2 - The statistical impact of sequencing and assembly metrics on the number of loci in which minor allele variation was detected 141

Figure 5.3 - Topology of the CAMP0252/Cj0284c locus 144
Figure 5.4 - Topology of the CAMP0410/Cj0448c locus 145
Figure 5.5 - Topology of Tlp8 (CetZ) encoded by the CAMP1031/Cj1110c locus with SNP data from ST-21 and Non ST-21 isolates. 145
Figure 5.6 - Topology of Transducer-like protein 3 (CcmL) highlighted with positions of variable residues as provided in Table 5.2 146
Figure 5.7 - Topology of the CAMP0922/Cj0999c locus encoding a putative integral membrane protein 147
Figure 5.8 – Topology of the CAMP1055/Cj1134 locus, encoding HtrB - an acyltransferase involved in lipid A biosynthesis 148
Figure 5.9 - Topology of the CAMP0303/Cj0337c locus encoding the flagellar motor proton channel protein, MotA 148
Figure 5.10 –Topology and arrangement of *C. jejuni*'s flagellar motor apparatus 149
Figure 5.11 - Topology of the CAMP1213/Cj1295 locus 149
Figure 5.12 - Circular chromosome of NCTC11168 with annotated CDS* and chromosomal positions of variable loci detected within the non ST-21 dataset 150
Figure 5.13 - Quality metrics of draft de novo assemblies, with mean inclusive of upper and lower confidence intervals highlighted in orange 151
Figure 5.14 - The statistical* impact of sequencing and assembly metrics on the number of loci in which minor allele variation was detected 152
Figure 5.15 - Topology of the CAMP1196/Cj1277c locus 154
Figure 5.16 - Topology of the *flaA* flagellin locus 155
Figure 5.17 - Circular chromosome of NCTC11168 with annotated CDS* and chromosomal positions of variable loci detected within the Poultry dataset 156
Figure 5.18 - Distribution of assembly metrics between the three deep sequencing datasets 157
Figure 5.19 - The statistical* impact of sequencing and assembly metrics on the number of loci in which minor allele variation was detected 158
Figure 5.20 - Circular chromosome of NCTC111168 with approximate* locations of variable loci detected across all deep sequencing isolates 160
Figure 5.21 – Loci containing synonymous and non-synonymous minor allele variants across all deep sequencing datasets. 162
Figure 5.22 – A representation of how haplotypes amongst a hypothetical *Campylobacter spp.* population may change as it moves from a chicken host to causing human infection 176

List of Tables

Table 1.1 - Species* of the *Campylobacter* genus and their major associated sources† 7

Table 1.2 - Source attribution of outbreaks reported to national surveillance schemes in the United Kingdom from 2010-2013 24

Table 1.3 - Reported place of campylobacteriosis acquisition amongst EU member states that provided data for all notified cases in 2013 25

Table 3.1 - MLST allele profiles for isolates with no identified sequence type 75

Table 3.2 Homology measured by wgMLST and mapping approaches of potentially linked *Campylobacter spp*. isolates 81

Table 3.3 - The position, alteration and effect of detected SNPs to their relative amino acid of a single-patient isolate pair 82

Table 4.1 - Patient demographics of samples belonging to the deep sequencing dataset . 109

Table 4.2 - Genes in which minor allele variation was detected, grouped according to putative function 111

Table 4.3 - Minor allele variants detected within loci encoding chemotaxis related proteins 113

Table 4.4 - Minor allele variants detected within loci encoding flagellar and membrane related proteins 117

Table 4.5 - Minor allele variation detected amongst multiple isolates in the Cj1321 homopolymeric tract 118

Table 4.6 - Minor allele variation present at intergenic positions 118

Table 4.7 - Loci in which the greatest frequency of unique alleles was observed amongst a group of isolates from the PubMLST database 119

Table 5.1 Patient demographics of samples belonging to the non ST-21 deep sequencing dataset 140

Table 5.2 - Genes in which minor allele variation was detected within non ST-21 dataset, grouped according to putative function 143

Table 5.3 - Sequencing coverage depth and clonal complex of the five isolates belonging to the poultry deep sequencing dataset 151

Table 5.4 - Genes in which minor allele variation was detected within poultry isolates, grouped according to putative function 153

Table 5.5 - Measures of the diversity observed in three deep sequencing datasets, normalised to account for the variance in sample size 159

Table 5.6 - –Locus tag*, function and known aliases of loci containing minor allele variation across all deep sequencing datasets 161

Table 9.1 - Strain and Genbank accession numbers for 16S rRNA genes used in the generation of Figure 1.2. 226

Table 9.2 - *Campylobacter spp.* incidence at Queen's Medical Centre, 2012-2014. 227

List of Abbreviations

ATP	Adenosine Triphosphate
CC	Clonal Complex
CDS	Coding Sequence
CFU	Colony Forming Units
CL2	Containment Level 2
DEFRA	Department for Environment, Food and Rural Affairs
DNA	Deoxyribonucleic Acid
FSA	Food Standards Agency
GBS	Guillain-Barré syndrome
HAP	Histidine Aspartate Phosphorelay
HK	Histidine Kinase
Hpt	Histidine Kinase Phosphotransfer
MA	Methyl-accepting
MAF	Minor Allele Frequency
MAMP	Microbe-associated Molecular Patterns
MAV	Minor Allele Variant
mCCDA	Modified Charcoal Cefoperazone Deoxycholate Agar
MCP	Methyl-accepting Chemotaxis Protein
MHB	Mueller Hinton Broth
MLST	Multilocus Sequence Typing
NGS	Next Generation Sequencing
PAS	Per-Arnt-Sim
PHE	Public Health England
QMC	Queen's Medical Centre
RNA	Ribonucleic Acid
SBS	Sequencing-by-Synthesis
SNP	Single Nucleotide Polymorphism
SOLiD	Sequencing-by-Oligo-Ligation-Detection
ST	Sequence Type
Tlp	Transducer like protein
wgMLST	Whole Genome Multi Locus Sequence Typing
WGS	Whole Genome Sequencing

Chapter One: Introduction

Campylobacter spp. are environmentally ubiquitous, and yet have paradoxically specific growth requirements with a notable sensitivity to heat, cold and ambient atmospheric conditions.

Despite this, *Campylobacter spp*. are highly successful, opportunistic human pathogens, and exist as the leading cause of bacterial enteritis in the developed world. In the EU, approximately 1 in 50 citizens will be affected by *Campylobacter* each year, at an estimated cost of €2.4 billion.

1.1. History

Campylobacter was originally known as a member of the *Vibrio* genus, a lack of suitable culture techniques limited comprehensive study of the organism and its impact on human health until the 1970's. However, the first notes on *Campylobacter* can be traced back to almost a century beforehand.

In 1886, Theodore Escherich published a series of articles in the German journal: Münchener Medizinische Wochenschrift (Escherich, 1886). Escherich discovered microscopic evidence of spiral shaped bacteria whilst examining colonic tissue from infants that died of cholera infantum; Escherich would also later find the same bacteria in the stool of children with gastroenteritis, but traditional culture methods failed to produce viable growth (Butzler 2004). The organisms were not aetiologically linked in either instance, and as such were largely ignored. It wasn't until a full century later that researcher M. Kist (1986), postulated that these articles likely described the earliest historic account of *Campylobacter* infection in humans.

Campylobacter was identified by veterinary researchers long before any links to human health were discovered; In 1906, McFadyen and Stockman described isolation of *Campylobacter* during an investigation into unusually high abortion rates amongst sheep flocks (McFadyen and Stockman, 1913). The researchers had expected to find a different organism: *Brucella abortus* – which happened to have similar growth

requirements to *Campylobacter*. At first they believed they had indeed found *B. abortus;* However, microscopic inspection revealed the same spirular bodies identified by Escherich, which at the time Fadyen classified as spirochetes (Skirrow 2006). Whilst conducting similar research into epizootic bovine abortions, Smith and Taylor again found the same 'vibrionic' bacteria, and were the first to formally designate them as *Vibrio fetus* (Smith & Taylor 1919).

More than a decade later, Jones *et al.* (1931), described another *Vibrio* species – this time found to be causing dysentery in calves. Isolated from the calf jejunum tissue, the new species was dubbed *Vibrio jejuni*. In 1938, *Vibro jejuni* was identified in the first well documented case of a campylobacteriosis outbreak in an American prison (Levy, 1946). 357 prisoners fell ill across two adjacent sites, with microscopy confirming the presence of *V. jejuni* in blood and stool samples. The source of the outbreak was attributed to an accidental delivery of raw milk in lieu of the institutions' regular pasteurised order - the supplier produced and bottled both products in identical, unlabelled 10-gallon jugs

In 1963, Sebald and Véron published a comprehensive study in the Annales de l'Institut Pasteur; their aim was to delineate members of the growing *Vibrio* genus. Chromatographic analysis of *Vibrio fetus* and *bubulus* yielded a GC content of 30-34% - far lower than other species - leading to the creation of a new genus: *Campylobacter*. From the Greek words *kampylos* and *bact* – meaning *curved* and *rod* respectively (Joshi, 2014.). *Campylobacter spp*. would be expanded in 1973 by Véron and Chatelain to include *C. jejuni, C. coli and C. sputorum*, beating other researcher's classification schemes to be accepted in the Approved List of Bacterial Names (1980).

Perhaps the most important breakthrough in *Campylobacter* research came from the formulation of a selective supplement containing vancomycin, trimethroprim and polymixin B – which is still in use today (Skirrow 1977, Urumova *et al.* 2014, Cean *et al.* 2013). This supplement offered a great improvement over the arduous filtration methods it

replaced, allowing laboratories to routinely isolate *Campylobacter* with ease and igniting novel interest in research of the genus. In the years following the deployment of this supplement, 12 new species were discovered and many more papers published. Figure 1.1 shows the drastic increase in published research following the introduction of the supplement.

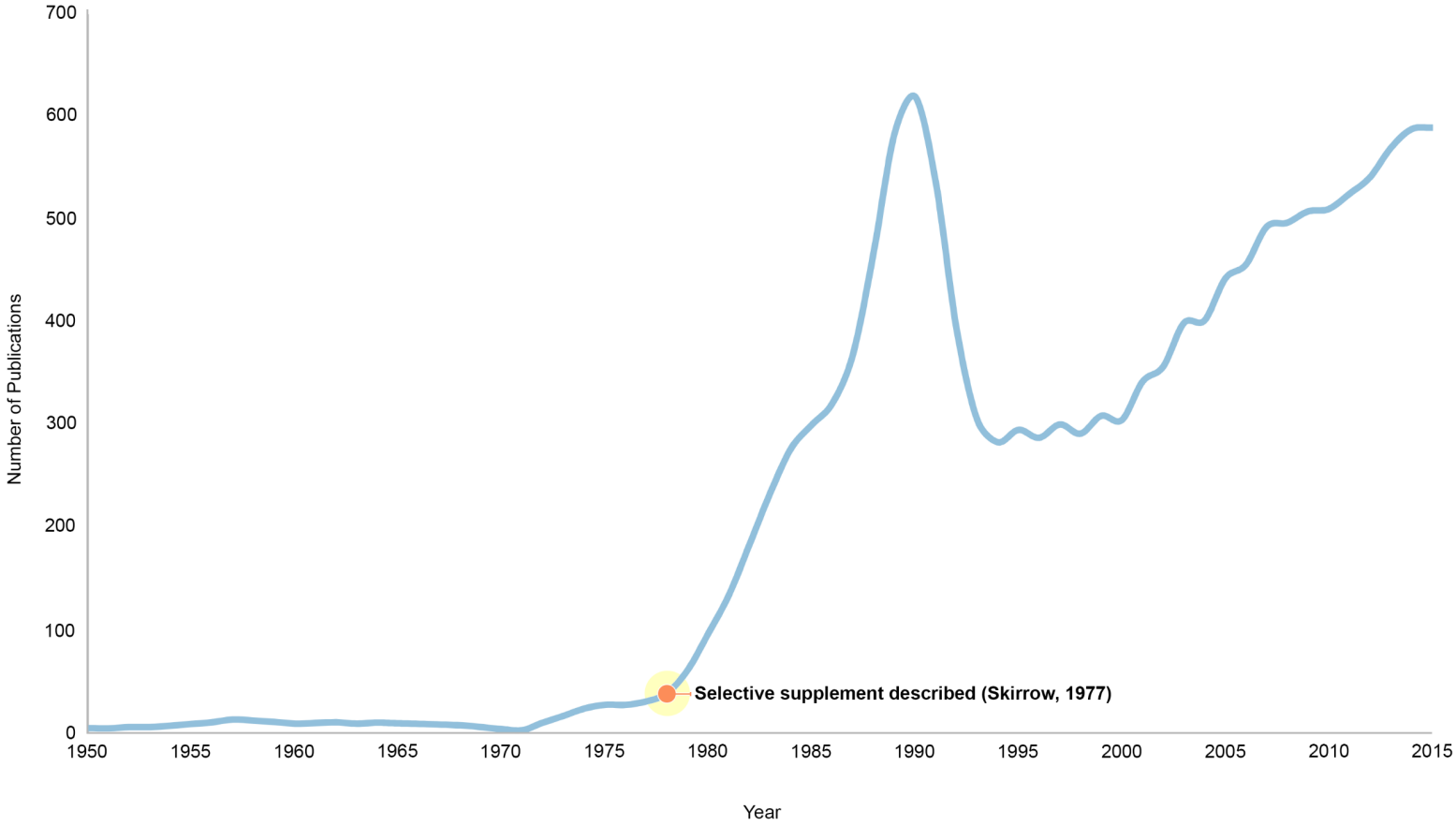

Figure 1.1 - Number of publications* with '*Campylobacter*' in the title before and after the introduction of Skirrow's Supplement

*Including only publications indexed by NCBI PubMed and PubMed Central databases.

1.2. Campylobacteraceae

Campylobacteraceae belong to the epsilon class of the *Proteobacteria,* and consists of three genera; *Campylobacter, Arcobacter* and *Sulfurospirillum* (Vandamme *et al.* 2008).

Since its initial description in 1963, the *Campylobacter* genus has undergone extensive change. Many organisms were spuriously included into the genus due to the biochemical heterogeneity of true *Campylobacter spp.*, which prevented a clear differentiation between *Campylobacter* and organisms with similar biochemical reactivity. As molecular techniques evolved, the genus was corrected and refined (Vandamme *et al.* 1991, Gebhart *et al.* 1985, Lawson *et al.* 1981, Kiehlbauch *et al.* 1991, Totten *et al.* 1985, Veron & Chatelain 1973, Tanner *et al.* 1981).

As of 2013, the *Campylobacter* genus contains 23 species and 4 subspecies, 18 of which have been associated with human disease (On 2013). Several sources cite a varying number of species within the genus (On 2001, Fitzgerald & Nachamkin 2011, Fernández *et al.* 2008). Even the most accurate species list will quickly become obsolete as novel *Campylobacter spp.* continue to be frequently isolated and described; *C. hepaticus, C. iguanorium* and *C. geochelonis* have all been proposed as potential additions to the genus over the past two years (Crawshaw *et al.* 2015, Van *et al.* 2016, Gilbert *et al.* 2015, Piccirillo *et al.* 2016).

Campylobacter spp. exhibit a high degree of variability in both phenotype and clinical relevance; Two thermotolerant species, *C. jejuni subsp. jejuni* and *C. coli,* are the main causes of human disease. *C. fetus* – the genus' type species – commonly causes epizootic abortion, though has in rare cases been linked to cases of human abortion (Lastovica & Allos 2008). Other notable species that are uncommonly involved in human pathogenesis include *C. upsaliensis, C. lari*, *C. hyointestinalis, C. jejuni subsp. doylei* and *C. sputorum* (Lastovica & Allos 2008), a list of

Campylobacter species and their associated source is provided in Table 1.2.

Table 1.1 – Species* of the *Campylobacter* genus and their major associated sources†

Taxon	Major Associated Host
Campylobacter avium	Poultry
Campylobacter canadensis	Whooping cranes
Campylobacter coli	Pigs, poultry, ostriches, ruminants
Campylobacter concisus	Humans, domestic pets
Campylobacter cuniculorum	Rabbits
Campylobacter curvus	Humans
Campylobacter fetus subsp. fetus	Cattle, sheep, reptiles
Campylobacter fetus subsp. venerealis	Cattle, sheep
Campylobacter gracilis	Humans
Campylobacter helveticus	Dogs, cats
Campylobacter hominis	Humans
Campylobacter hyointestinalis subsp. hyointestinalis	Cattle, deer, pigs, hamsters
Campylobacter hyointestinalis subsp. lawsonii	Pigs
Campylobacter insulaenigrae	Seals, porpoises
Campylobacter jejuni subsp. doylei	Humans
Campylobacter jejuni subsp. jejuni	Poultry, cattle, pigs, wild birds
Campylobacter lanienae	Cattle
Campylobacter lari subsp. concheus	Shellfish
Campylobacter lari subsp. lari	Wild birds, dogs, poultry, shellfish,
Campylobacter mucosalis	Pigs
Campylobacter peloridis	Shellfish
Campylobacter rectus	Humans
Campylobacter showae	Humans
Campylobacter sputorum	Humans, pigs, ruminants
Campylobacter subantarcticus	Birds in the subantarctic
Campylobacter upsaliensis	Dogs, cats
Campylobacter ureolyticus	Humans
Campylobacter volucris	Black-headed gulls

* - Taxa listed in bold have been shown to cause disease in humans.
† - Table adapted from (On, 2013).

The advent of next generation sequencing and progressive reduction in associated costs have allowed robust phylogenetic analyses, starting with

simpler methods based on one or many genes to whole-genome methods. The 16S rRNA gene is ubiquitous amongst prokaryotes and has a clock-like mutation rate. It contains both variable and conserved regions that aid in both target amplification and differentiation, making it an ideal candidate for amplicon sequencing (Mizrahi-Man *et al.* 2013). The level of variation present across the gene's nine assigned (V1-9) regions is sufficient to delineate taxa and can provide a basic level of phylogeny. Figure 1.2 shows the maximum likelihood phylogeny of all *Campylobacter spp*. 16S regions available at the time of analysis.

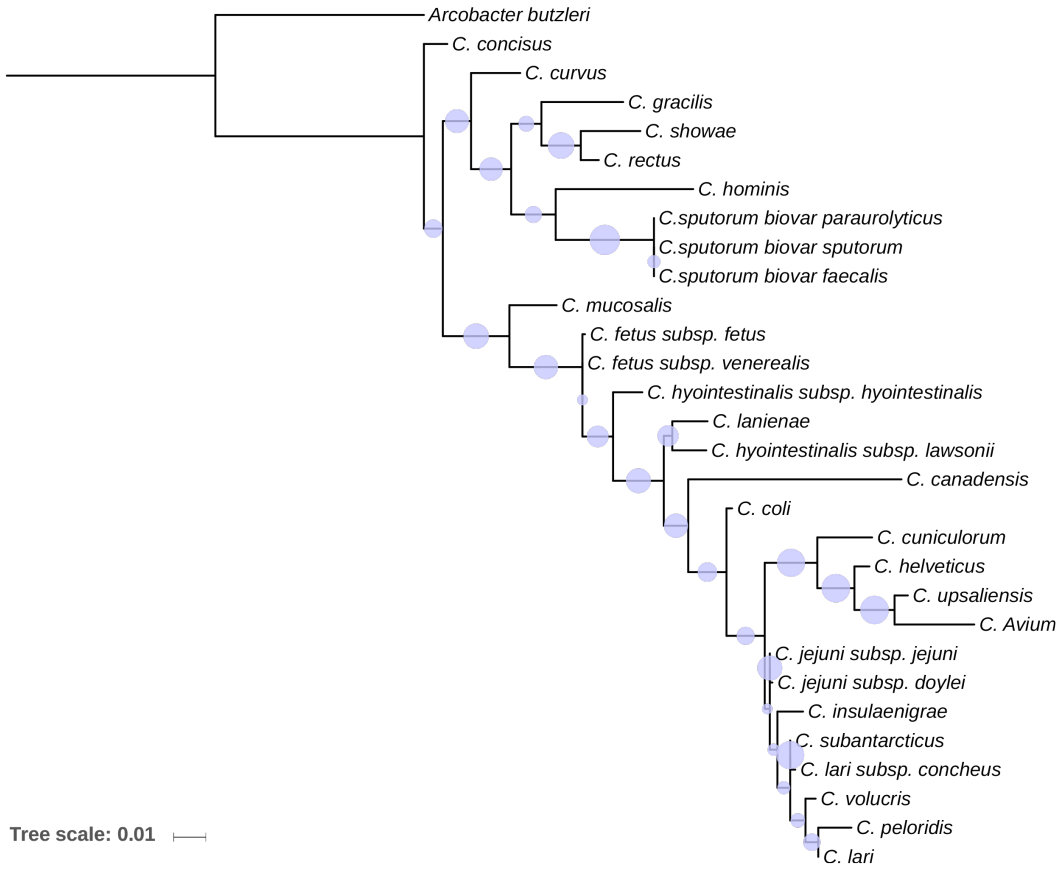

Figure 1.2 - Phylogenetic analysis of *Campylobacter* 16S rRNA regions

The evolutionary history of the 16S rRNA gene was inferred using the maximum likelihood method (Tamura & Nei 1993). The tree with the highest log likelihood (-6409.8617) is shown. The tree is rooted to *Campylobacteraceae* family member *Arcobacter butzleri*, with bootstrap values represented by diameter-relative circles.
Accession numbers of the gene sequences used in this analysis provided in Appendix 9.1

The *Campylobacter* genus consists of fastidious, non-sporulating, Gram negative bacteria. They are relatively small organisms (0.2-0.8 µm wide x 0.5-5 µm long), which under optimal growth conditions exist as spiral or

'S' shaped rods (Vandamme *et al.* 2008, Silva *et al.* 2011). *Campylobacter spp.* are capnophillic and thrive in atmospheres containing between 3-5% O_2, and 3-15% CO_2.

Campylobacter have single or dipolar unsheathed flagella which can be 2-3x larger than the length of the cell (Lous *et al.* 2011, Vandamme *et al.* 2008). This - along with their helical morphology - generates *Campylobacter*'s characteristic corkscrew motility, conferring the ability to attach and invade to intestinal epithelial cells (Lugert *et al.* 2015). The only exceptions to this are *C. gracilis* – a non-motile *Campylobacter spp.* – and *C. showae,* which has multiple flagellar (Etoh *et al.* 1993, Vandamme *et al.* 1995). The majority of species reduce fumarate to succinate, reduce nitrate, yield a negative methyl red reaction, produce indole and acetoin and (excluding *C. jejuni*) do not hydrolyse hippurate (Garénaux 2008).

Campylobacter spp. are also pleomorphic; as cells approach stationary phase – or under stress conditions – they take on a coccoid form. Whilst the cells still retain their flagellar in this configuration, they lose their motility and thus their ability to invade host cells (Ikeda & Karlyshev 2012). This form has also been linked to the production and maintenance of biofilms (Ikeda & Karlyshev 2012, Walker *et al.* 1986).

1.2.1. Thermotolerant Species

Camypylobacter jejuni subsp. jejuni – hereon referred to as *C. jejuni* unless specified otherwise – and *Campylobacter coli* are responsible for the majority of human disease. These organisms preferentially grow between 37 and 42°C and as such are referred to as thermotolerant species (Levin 2007, Silva *et al.* 2011). Whilst other species (*e.g. C. lari, C. upsaliensis*) are also thermotolerant, this chapter will focus on *C. jejuni* and *C. coli* specifically.

Thermotolerant *Campylobacter spp.* are microaerophillic organisms, with an atmospheric optima of 5% O_2, 10% CO_2 and 85% N_2. Under ambient atmopheric conditions, an accumulation of cellular peroxides -

mostly reactive oxygen species (ROS) - inhibits growth and proliferation (Garénaux 2008). *In vitro* studies have shown exposure to ambient atmosphere and temperature results in a significant loss of viability from as little as 10 minutes (Davis & DiRita 2008).

Campylobacter spp. are fastidious, non-fermentative organisms. *C. jejuni* lacks the ability to catabolise glucose due to the absence of 6-phospofructokinase, and instead relies on citric acid cycle intermediates and a handful of amino acids acquired in the gut (Kelly 2001, J Parkhill *et al*. 2000, Stahl *et al*. 2012). Typically *C. jejuni* utilises amino acids sequentially: serine, aspartate, asparagine and glutamate, in order of preference. Many of these amino acids are major components of chicken excreta (Garénaux 2008).

During the process of colonisation, *Campylobacter* must compete with the natural intestinal flora for nutrients. Iron catalyses several biochemical reactions and is an essential component in microbial metabolism; free iron availability in host-tissues is limited, with most iron bound in protein complexes such as haemaglobin (Naikare *et al*. 2013). *C. jejuni* has multiple siderophore mediated iron uptake systems that allow it to access sufficient iron, including enterobactin, haem, transferrin and rodotorulic acid (Miller *et al*. 2009, Naikare *et al*. 2013).

1.2.1.1. Isolation

Under optimal conditions, *Campylobacter spp*. can produce growth on basal media, however in food and clinical laboratories it is generally necessary to use a supplemented media to provide a sufficiently pure and substantial recovery (Garénaux 2008).

Historically, isolation of *Campylobacter spp*. would require a round of membrane filtration followed by growth on a blood-based agar. The addition of blood serves to quench toxic oxygen metabolites and subsequently reduce the level of oxidative stress exerted on *Campylobacter spp*. (Corry, Atabay, Forsythe, & Mansfield, 2003). Supplementing agar with blood provides a challenge in itself, as the

laboratory has to ensure that sufficiently fresh and sterile stocks are constantly available.

Modern *Campylobacter* media replaces the oxygen quenching action of blood with anhydrous chemical alternatives (*e.g.* ferrous sulphate, sodium pyruvate), which greatly simplifies the media preparation process. The International Organisation for Standardardisation methodology (ISO 10272) for isolating *Campylobacter spp.* from food and clinical samples recommends the use of 'Modified Cefaperazone Charcoal Deoxycholate Agar' (mCCDA) . This agar uses Nutrient Broth as a base and contains charcoal, casein, sodium pyruvate, sodium deoxycholate and ferrous sulphite. The media is supplemented with Cefaperazone (a cephalosporin that thermophilic *Campylobacter* is resistant to) and Amphotericin B to inhibit bacterial and fungal growth respectively.

A pre-enrichment stage can also be used to increase the recovery of sub-lethally injured cells; liquid culture in a suitable broth (*e.g.* Müeller-Hinton) prior to plating on mCCDA has been shown to increase the consistency of isolation from several sample types, including food, faeces and water (Jokinen *et al.* 2012, Humphrey 1989).

1.2.1.2. Identification

The appearance of *Campylobacter spp.* can vary depending on the medium used. Typically, *Campylobacter* produces a flat to slightly convex, irregular, effuse growth and swarming is observed if the media is sufficiently moist (Garénaux, 2008; Skirrow & Benjamin, 1980). Initially, colonies are transparent in appearance, transitioning to a metallic grey as the culture matures (Garénaux, 2008; Skirrow & Benjamin, 1980).

In addition to a Gram stain, suspected colonies should be further identified via biochemical tests. Several clinically relevant *Campylobacter spp.* are oxidase positive (Garénaux 2008). *C. jejuni* can be differentiated from other species using a hippurate hydrolysis test

(positive). Colonies that produce typical *Campylobacter* growth characteristics, a negative Gram stain and a positive oxidase result can be clinically reported as *Campylobacter* for diagnostic purposes (Barrow, 2003; MacFaddin, 1980).

To remove any degree of ambiguity, a definitive and rapid method for identification and speciation of *Campylobacter spp*. is PCR. Several schemes have been developed, based on amplicons of the 16S rRNA sequence, *hipO* hippuricase and *ipxA* acyltransferase genes (Linton *et al*. 1997, Klena *et al*. 2004, Linton *et al*. 1996).

1.3. Clinical manifestations of *Campylobacter*

Infection with *Campylobacter spp*. is known under the generic description campylobacteriosis (Coker *et al*. 2002). Due to the overwhelming proportion of cases that manifest as gastroenteritis, the term campylobacteriosis has become widely synonymous with *Campylobacter* enteritis and will be referred to as such hereafter (Georg Kapperud *et al*. 1992, Bokkenheuser *et al*. 1979, Guerrant *et al*. 1978, Eberhart-Phillips *et al*. 1997, Mossong *et al*. 2016, Pearson *et al*. 2000, Frost 2001, Wilson *et al*. 2008, Wingstrand *et al*. 2006, Olson *et al*. 2008).

1.3.1. Campylobacteriosis

Campylobacter jejuni is widely accepted as the main causative agent of campylobacteriosis and is responsible for ~80-90% of all reported cases, with *C. coli* implicated in ~7-12% (Gillespie *et al*. 2002, Chen *et al*. 2011, Olson *et al*. 2008). Both species produce identical clinical presentation, which in itself is indistinguishable from enteritis mediated by other enteropathogens (Acheson & Allos 2001).

Campylobacter has a low infectious dose; It has been previously demonstrated in volunteer studies that an intestinal infection can be propagated following ingestion of just 800 organisms (Black *et al*. 1988). Similarly, Robinson (1981) reported a dysenteric episode following self-

administration of an inoculum containing 500 organisms. The infectious dose is much lower than other common enteropathogens, for example *Salmonella spp*. and some strains of *E. coli* which have infectious doses in the 10^5 – 10^9 range (Kothary & Babu 2001).

Following ingestion, *Campylobacter* undergoes an incubation stage typically lasting for 1-7 days - notably longer than other intestinal infections (Silva *et al*. 2011). On average this period will last 72 hours, though periods in excess of 7 days have also been reported, which may be linked to the consumption of fewer viable organisms (Janssen *et al*. 2008, Silva *et al*. 2011, Walker *et al*. 1986).

In 30% of campylobacteriosis cases, the diarrhoeal phase of the illness is preceded by a prodromal stage (Blaser & Engberg 2008). Prodromal patients may present in a flu-like condition up to 24-48 hours before developing diarrhoea, with symptoms including pyrexia, myalgia, malaise and abdominal pain. The occurrence of this prodrome prior to gastrointestinal symptoms indicates that the initial inflammatory response to the infection occurs before intestinal epithelial cell function is diminished (Wassenaar & Blaser 1999).

Campylobacteriosis is characterised by an inflammatory dysenteric episode which generally lasts between 2 and 7 days. At the peak of the diarrhoeal stage, patients will pass 8-10 stools per day; stools are profuse and watery, and in ~15-30% of instances contain visible frank blood (Blaser *et al*. 1983, Gillespie *et al*. 2006). Irrespective of blood visibility, faecal leucocytes and erythrocytes are present in the majority of cases indicating colonic and rectal tissue invasion (Blaser *et al*. 1979). Abdominal pain is reported in >85% of cases and can mimic appendicitis due to its severity and duration. Vomiting is reported in ~35% of cases and is generally associated with a longer illness and greater likelihood of hospital admission (Gillespie *et al*. 2006).

In the majority of cases, the disease is self-limiting with a gradual relief of symptoms over several days, and at most may only require supportive

treatment (*i.e.* isotonic hydration). However, the severity of the disease varies greatly, with some patients exhibiting a mild, short term diarrhoea and others experiencing far more severe and enduring symptoms. This is likely a multifactorial effect due to strain pathogenicity, number of viable organisms consumed and host immune response (Silva *et al.* 2011).

In more severe cases of campylobacteriosis, where patients exhibit a high fever, excessively bloody stools or have a prolonged (*i.e.* >7 days) illness, antibiotic treatment should be administered. Previously *Campylobacter* infections were treated with fluoroquinolones, however a global increase of fluoroquinolone resistance amongst *Campylobacters* mean physicians now typically administer macrolides such as erythromycin or clarithromycin (Cody *et al.* 2012).

With the exception of patients treated with antibiotics, *Campylobacter* will continue to be faecally shed for several weeks following symptom relief, with a mean excretion period of 37.6 days (Blaser, MJ; Engberg 2008). Whilst transmission from convalescent excretion of *Campylobacter* has yet to be demonstrated, relapses have been reported to occur in 15-25% of cases (Kapperud *et al.* 1992; Blaser & Engberg 2008).

1.3.2. Extraintestinal Complications

Whilst uncommon, infection with *Campylobacter* can be complicated by extraintestinal localisation. Case studies reveal instances of campylobacteriosis resulting in cholecystitis, pancreatitis, nephritis, myocarditis and hepatitis (Korman *et al.* 1997, Cunningham *et al.* 2003, Op den Winkel *et al.* 2010, Udayakumar & Sanaullah 2009).

Bacteraemia is a more common – albeit still relatively rare – complication, with an incidence of 1.5-8 in 1,000 cases. Bacteraemia occurs predominantly in elderly and immunocompromised (*e.g* HIV positive) cohorts, with the latter group often exhibiting no gastrointestinal symptoms (Skirrow *et al.* 1993, Fernández-Cruz *et al.*

2010, Georg Kapperud *et al.* 1992). It is postulated that transient bacteraemia may occur more frequently than the reported incidence would suggest, particularly in febrile patients. The true occurrence is difficult to determine as blood cultures are rarely obtained from patients with acute gastroenteritis and detection methods are not necessarily sensitive to *Campylobacter* species, which are inherently weak to serum (Tee & Mijch 1998, Pacanowski *et al.* 2008, Shandera *et al.* 1992).

1.3.3. Guillain-Barré Syndrome

Guillain-Barré syndrome (GBS) is the most common cause of acute flaccid paralysis in countries where poliomyelitis has been eradicated, with annual incidence ranging from 0.6-4/100,000 people (Shahrizaila & Yuki 2016). It is characterised by a symmetrical ascending weakness that develops over the course of 2-3 weeks; most patients present with hypoflexia and weakness in the legs or arms. The severity of the paralysis varies, with some patients requiring ventilation within days and others not experiencing any loss of ambulation (Dimachkie & Barohn 2013).

In 40-70% of cases, GBS is preceded by an antecedent infection (McGrogan *et al.* 2009). In ~30% of cases that antecedent is *C. jejuni,* which is the most common aetiological agent of GBS. It is estimated that ~0.1% of campylobacteriosis cases will develop into GBS (Nachamkin *et al.* 1998). It is hypothesised that campylobacteriosis propagates the development of antibodies to lipooligosaccharides which exhibit cross reactivity with peripheral nerve gangliosides. This is supported by research that demonstrates antiganglioside antibody depletion by *C. jejuni* LOS, and anti-LOS antibody depletion by gangliosides (Ang *et al.* 2002, Godschalk *et al.* 2004).

1.3.4. Transmission

Campylobacter spp. can be isolated from a wide range of host species, including cattle, sheep, pigs and birds (Silva *et al.* 2011). There is evidence that suggests *Campylobacter* acts as a commensal organism in some hosts, and acts as an opportunistic pathogen in others (Waldenström *et al.* 2002, French *et al.* 2009, Açik & Çetinkaya 2006,

Stanley & Jones 2003, Wesley *et al.* 2000, Young *et al.* 2000, Padungton & Kaneene 2003).

Human infections with *Campylobacter spp.* are considered to be food-borne, with contaminated meat or other animal-sourced products (*e.g.* milk) responsible for the majority of cases. In general, during the slaughter process intestinal contents and bacteria on the skin surface of the animal come into contact with the meat, equipment, surfaces and process water, with contamination persisting throughout the processing chain to the end consumer. Improper handling or consumption of undercooked meat products provides an ample opportunity for infection to develop (Skarp *et al.* 2016).

1.3.4.1. Poultry

Poultry is the main food source – and therefore largest single source - of human campylobacteriosis (EFSA 2014, Skarp *et al.* 2016). Chicken (*Gallus gallus)* constitutes 70-80% of poultry meat production (Skarp *et al.* 2016), with turkey and duck meat also included in this category.

Due to the sporadic nature of *Campylobacter* infections, the exact attribution rate is difficult to determine, however several epidemiological studies have shown a significant association between handling or consuming raw or improperly prepared poultry and contraction of *Campylobacter* enteritis (Vellinga & Van Loock 2002, Stern *et al.* 2003). In 1999, Belgium had to withdraw chicken meat from sale due to contamination with dioxin. During the withdrawal period there was a 40% reduction in campylobacteriosis cases (Vellinga & Van Loock 2002). Similarly, in 1996 Icelandic chicken sales shifted largely from frozen to fresh cuts, which also coincided with a large peak campylobacteriosis incidence (Stern *et al.* 2003).

Campylobacter has specific molecular tropisms that favour the colonisation of the avian gut; its preferred nutrient sources are abundant in chicken ceca and it thrives in the warm (42°C), microaerobic atmosphere of lower intestinal tract (Newell & Fearnley 2003).

Colonisation in chickens has been experimentally achieved with a dose as low as 40 CFU (Cawthraw *et al.* 1996), and in one instance, 8 CFU (Hendrixson *et al.* 2001). Once established, *Campylobacter*s are rapidly detected in excreta, and reach as many as 10^9 CFU in the caecal contents (Hermans *et al.* 2011, Stern *et al.* 2001). The faecal shedding of *Campylobacter* and coprophagic nature of chickens leads to a rapid dissemination amongst commercial flocks, with whole-flock positivity occurring within 7 days of the initial point-zero colonisation (Hermans *et al.* 2012). *Campylobacter* is therefore understandably profuse amongst commercial flocks; 75-90% of all British flocks sampled are found to be colonised (FSA & BBSRC 2013, EFSA 2014, Wilson *et al.* 2008)

At slaughter, *Campylobacter* contaminates meat via the spillage of intestinal contents. *Campylobacter* can also be found on the feathers, in the air of 'unclean zones' (*e.g.* hanging room, eviscerating room) and on processing equipment across the production line (Berndtson *et al.* 1996). EFSA estimate that decreasing the *Campylobacter* load on poultry products by 1 log could reduce the risk of human disease by 50-90% (EFSA 2011a). As a consequence, several biosecurity measures at the farm level and subsequent interventions at slaughter have attempted to achieve this, with varying levels of success (EFSA 2011; Food Standards Agency 2016).

Despite significant efforts, the level of *Campylobacter* that persists in the processing chain to retail remains high. In 2014, an average of 71% of retail chicken was found to be contaminated with *Campylobacter* - 21.8% of those samples were shown to harbour in excess of 1000 CFU/g (Food Standards Agency 2016).

The massive prevalence of *Campylobacter* enteritis can in part be explained by examining consumer eating habits. The British public consume over 30 kg of poultry per capita each year, the highest of all European countries (FAO 2016). In 2014, the UK spent £4.29 billion on chicken alone, and consumption is continually increasing (Key Note 2015). There is also a strong appetite for poultry meat on a global scale;

Saint Vincent for example consumes 74 kg per capita per year. Global consumption is rising independently of population growth by approximately 2.6% annually (FAO 2016).

Poultry is an inexpensive protein source, and is consistently one of the cheapest meats available in terms of cost per kilogram. In a period covering 1987-2014, poultry prices rose by just ~45% compared to beef, lamb and pork at ~120%, ~190% and ~135% respectively (Key Note 2015). This, combined with the versatility and familiarity of poultry means household reliance and consumer preference is sustained.

1.3.4.2. Livestock

Other food-animals are also a potential source of *Campylobacter* infection. Beef – particularly beef liver – has been shown to harbour *Campylobacter* (Osano & Arimi 1999, Noormohamed & Fakhr 2013), with 38% of bovine faeces containing *C. jejuni* (Inglis *et al.* 2003). Approximately 10% of clinical campylobacteriosis is attributable to beef (Dearlove *et al.* 2015). Pork is generally associated with a higher prevalence of *C. coli* compared to *C. jejuni,* though the reservoir as a whole is only estimated to be responsible for 1% of campylobacteriosis in the UK (Dearlove *et al.* 2015).

1.3.4.3. Companion animals

Campylobacter can be isolated from household animals which may not display typical campylobacteriosis symptoms. Dogs and cats – particularly puppies and kittens – have often been shown to harbour several *Campylobacter spp.,* the most common being *C. upsaliensis* (an emerging human pathogen), *C. jejuni* and *C. coli* (Gras *et al.* 2013, Skirrow 1977, Baker *et al.* 1999). Whilst pet ownership increases the risk of contracting campylobacteriosis, basic hygiene methods prevent any significant transmission in all but the youngest groups (0-2 years).

1.3.4.4. Milk

Raw milk has historically been implicated in a number of *Campylobacter* outbreaks (Fahey *et al.* 1995, Porter & Reid 1980, Wood 1992). Whilst the increased use of pasteurisation has largely eliminated such occurrences, failure in the pasteurisation process, cross contamination or indeed continued reliance on raw milk products still pose a threat to public health, particularly in developing countries (Heuvelink *et al.* 2009, LeJeune & Rajala-Schultz 2009, Fahey *et al.* 1995).

1.3.4.5. Other sources

The number and variety of *Campylobacter* sources is vast, though all share the commonality of faecal contamination. The environmental ubiquity and diverse commensality of *Campylobacter spp.* provides a multitude of potential vectors.

Whilst uncommon, direct contact with domesticated animals shedding *Campylobacter* is a noted risk factor in disease transmission, particularly amongst veterinary workers (*e.g.* farmers, veterinarians) (On 2001).

Water has been attributed to some incidences of *Campylobacter* transmission. Drinking from or swimming in lakes or diving in waters approximate to sewage outflow pipes are both noted examples (Obiri-Danso & Jones 1999, Jones 2001). In the developed world, sanitary treatment eradicates drinking water as a potential source of *Campylobacter* acquisition, however accidental consumption of untreated water has historically been linked to outbreaks (Vogt *et al.* 1982, Broczyk *et al.* 1987). In developing countries where treated water is scarce, water sources are still frequently identified as a major environmental source (Ashbolt 2004).

Given the frequent isolation of *Campylobacters* from wild birds (Yogasundram *et al.* 2013, Waldenström *et al.* 2010), faecal contamination of water is common; whilst this commonality does not directly result in a significant number of transmission events it does

serve to spread the organism to other hosts that bathe in or drink from the water source. Muscid flies (*e.g. Musca domestica*) have been proposed as potential mechanical vectors (Ekdahl *et al.* 2005) that serve to spread *Campylobacter* to broiler flocks, work surfaces, and water bodies.

1.3.5. Incidence and Prevalence

Campylobacter is the leading cause of bacterial enteritis in the developed world. In the United Kingdom, it is responsible for approximately 22,000 hospitalisations and more than 100 deaths each year; it is also a significant burden to the UK economy, costing tax payers £900 million per annum (DEFRA & FSA, 2013).

In 2013, Public Health England received 66,575 reports of campylobacteriosis (DEFRA 2015). Whilst this is 11.1% higher than the 10-year average (~60,593) it is an 8.2% reduction from 2012's incidence (72,592), which holds the highest number of cases recorded since the surveillance scheme was introduced (PHE & DEFRA 2013, DEFRA 2015, DEFRA & PHE 2015). Preliminary data from England and Wales suggests that incidence was largely unchanged in 2014 and 2015 (Public Health England, 2015; 2016).

1.3.5.1. Under-reporting

Campylobacteriosis is a severely under-reported condition. This may be in part due to milder manifestations of the disease, both in terms of duration and symptom severity. In several documented outbreaks the percentage of infected persons who sought medical attention ranges from 1-50% (Pearson *et al.* 2000, Inns *et al.* 2010, Korlath *et al.* 1985, Frost 2001, Fahey *et al.* 1995); in sporadic cases – which constitute the majority of *Campylobacter* infections - it is much harder to extrapolate.

Several studies have tried to calculate the degree of under-reporting, which varies greatly across the world. In the United Kingdom it is estimated that only 1 in 9.3 cases are reported to the relevant authorities (Tam *et al.* 2012).This means in 2013, the total number of *Campylobacter* infections across the UK was closer to ~620,000.

Similar data from the Netherlands and the United States of America show a reporting ratio of 1:12.1 and 1:30.3 respectively (Scallan *et al.* 2011, Havelaar *et al.* 2012). Figure 1.3 shows estimated incidence range of *Campylobacter* across 28 European Union member states based on individually calculated reporting ratios and illustrates how variable *Campylobacter* incidence can be; adjacent states display a largely different incidence due to various socioeconomic and dietary factors, as well as differences in pathogen surveillance schemes.

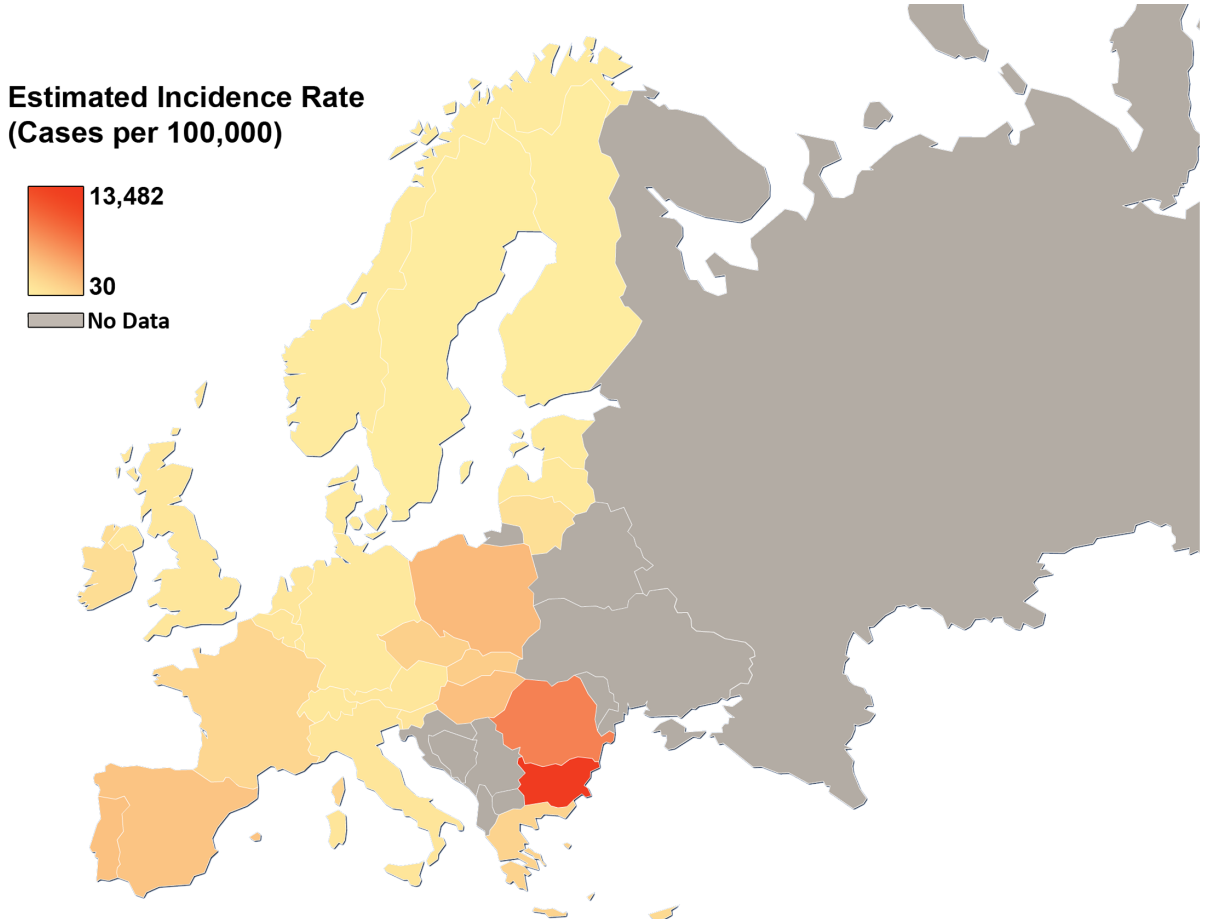

Figure 1.3 - Incidence rates of *Campylobacter* infections in European countries based on individually calculated reporting ratios*

*Data obtained from the EFSA (2014), accounts for underreporting of community based infections.

1.3.5.2. Case Linkage

Infections with *Campylobacter spp*. are highly sporadic, with very few cases of epidemiologically related cases identified outside of notified outbreaks (Blaser 1997, Olson *et al*. 2008). This is in part due to the vast genomic diversity of *Campylobacter spp.*, and the variances in symptom severity and by extension, reporting frequency.

1.3.5.3. Outbreaks

In 2010, *Campylobacter* displaced *Salmonella* as the leading cause of foodborne infectious disease outbreaks in England and Wales, with 21 occurrences accounting for 18.3% of reported outbreaks. Table 1.2 shows the number of outbreaks and their resolved source in the UK from 2010-2013; the majority (79.1%) of outbreaks were traced to poultry meat, with chicken liver pâté implicated in particular across numerous cases.

A similar number of outbreaks occur annually in other parts of the world; In the USA, there are on average 28 outbreaks of campylobacteriosis per year, with dairy products identified as the most common source (Taylor *et al*. 2012). In 2013, the EFSA received 414 reports of *Campylobacter* outbreaks across 16 European Union member states, which accounted for 8% of all food-borne outbreaks that year (EFSA 2014).

Table 1.2 - Source attribution of outbreaks reported to national surveillance schemes in the United Kingdom from 2010-2013

Source	Number of Outbreaks/Year				% of Cases
	2010	2011	2012	2013	
Poultry meat	16	17	6	14	79.1%
Red meat	2	0	1	1	6.0%
Composite foods	2	0	0	0	3.0%
Milk/Dairy	0	0	0	1	1.5%
Potable water	1	0	0	0	1.5%
Unidentified	0	3	0	3	9.0%
Total outbreaks	21	20	7	19	

Table compiled from data provided by Public Health England and the Department for Environment, Food and Rural Affairs (Health Protection Report, 2010; Health Protection Report, 2011; Zoonoses Report, 2012; Zoonoses Report, 2013).

1.3.5.4. Developing countries

In developing and tropical countries, *Campylobacter* is hyperendemic and presents a notably different epidemiology. Incidence amongst children under 5 years of age is significantly higher, with estimates ranging from 40,000 to 60,000/100,000, compared to 300/100,000 in the USA and other industrialised nations (Coker *et al.* 2002, Kotloff *et al.* 2013, Lee *et al.* 2013, Janssen *et al.* 2008).

Infection rates peak in childhood and decline with age, in contrast to the United Kingdom where infection rates peak in the over 60 group(FSA & DEFRA 2010). Additionally, asymptomatic carriage is more common in regions where *Campylobacter* infection is endemic. Conversely, outbreaks are far less common. It's postulated that these differences may be a result of developing immunity due to frequent exposure (Havelaar *et al.* 2012, Blaser *et al.* 1985).

1.3.5.5. Traveller's Diarrhoea

The risk of contracting diarrhoea whilst travelling is approximately 7% in developed nations, and ranges from between 20-50% in the developing world (Ardley & Wright 2010). The majority of *Campylobacter* infections are not formally investigated in the United Kingdom and data for whether *Campylobacter* was contracted internationally or domestically is limited.

In 2013, the UK reported 66,465 cases of *Campylobacter* infections to the EFSA - 15.9% of those cases had place-of-acquisition data. Of this subset, ~6.3% of campylobacteriosis cases were travel-related (EFSA 2014). A total of 7 EU member states report on the travel/domestic status of all notified *Campylobacter* infections (Table 1.3); on average 4.27% are acquired internationally, which validates the extrapolation of UK data.

Table 1.3 - Reported place of campylobacteriosis acquisition amongst EU member states that provided data for all notified cases in 2013

Country	No. cases	International	Domestic
Austria	5,726	9.3%	90.7%
Czech Republic	18,267	1.6%	98.4%
Estonia	382	4.7%	95.3%
Germany	63,271	8.2%	91.8%
Netherlands	3,702	5.5%	94.5%
Poland	552	0.2%	99.8%
Slovakia	5,845	0.4%	99.6%
EU Total	**97,745**	**4.27%**	**95.73%**

Table compiled from data provided by EFSA, exclusively including states that had 0% of cases demarked as undetermined and more than 100 reported cases (EFSA 2014).

1.3.5.6. Epidemiology

Campylobacter epidemiology has a bimodal distribution, mainly affecting children under 4 years of age, and adults over 60 (FSA & DEFRA 2010). The high prevalence amongst young children mimics the epidemiology of other enteric pathogens (*e.g. Yersinia enterolitica*) and may be due to a number of factors (Rosner *et al.* 2010, Tam *et al.* 2012). Children generally have a lower level of hygiene, and their immune system is still naïve and underdeveloped. Additionally, a higher portion of cases will likely be reported than in other age groups due to parental concern (Gillespie *et al.* 2009, Schielke *et al.* 2014). In recent years, incidence amongst elderly cohorts has increased, this is thought to be due to the general ageing of the underlying population base (Nichols *et al.* 2012).

The number of *Campylobacter* cases reported per year is higher amongst males than females. This increased incidence is observed across all age groups at a 1.14 M/F ratio, suggesting an inherently greater susceptibility rather than involvement in risk related behaviours (Nichols *et al.* 2012, Louis *et al.* 2005).

1.3.5.7. Seasonality

Campylobacter displays a distinct seasonality, with infection rates soaring during the summer months. Whilst incidence varies by region, a peak from weeks 18-22 (May-June) is consistently observed across the entirety of the UK, with the highest number of cases occurring in June (Nichols *et al.* 2012, Louis *et al.* 2005). The forces driving *Campylobacter* seasonality are ill defined, but many hypotheses have been presented, including agricultural activities, weather, increased poultry contamination and an increase in the population of flies (Ekdahl *et al.* 2005, Nichols *et al.* 2012, Louis *et al.* 2005).

1.3.5.8. High Risk Groups

Elderly and immunocompromised groups are disposed to a higher risk of contracting campylobacteriosis; *C. jejuni* incidence is 40-100X higher amongst AIDS afflicted persons (Sorvillo *et al.* 1991, Larsen *et al.* 2011). These groups are also at greater risk of experiencing complications and as such have a higher mortality rate than other cohorts.

Gastric acidity is one of the first defensive barriers to infection, and is particularly effective against *Campylobacter*. However, patients that are taking proton pump inhibitors – for example Omeprazole, which is now available over the counter – are at greater risk of contracting campylobacteriosis (Neal *et al.* 1996).

1.3.6. Pathogenesis

Despite a substantial amount of research, the mechanisms underpinning *Campylobacter* pathogenesis are still poorly understood. Following ingestion of a sufficient quantity of bacteria, *Campylobacter* must survive the host environment and migrate to their target organ, the small intestine. *Campylobacter* utilises its motility and chemotactic apparatus to reach its primary destination - the intestinal mucus layer - which serves as a base for host-cell adhesion and invasion.

1.3.6.1. Motility

Campylobacter exhibits a corkscrew-like pattern of motility conferred by a single flagellum at one or both poles. The flagellum filament is composed of a flagellin subunit polymer and encoded by the *flaA* and *flaB* genes. Mutagenesis studies have shown that *flaA* knockout produces a truncated flagella that completely disrupts motility, whereas *flaB* knockout only yields a slight decrease, indicating that *flaA* is the major flagellin component (Wassenaar *et al.* 1991, Wassenaar *et al.* 1993, Wassenaar *et al.* 1994). The larger flagella complex contains multiple units, including a motor, basal body, and stator ring complex.

Campylobacter has a particularly powerful motor complex that allows it to bore through intestinal mucus and between epithelial tight-junctions (Beeby *et al.* 2016).

Campylobacter's flagella is unusual in that it lacks several microbe-associated molecular patterns (MAMPs) that are highly conserved amongst other bacterial flagellins (De Zoete *et al.* 2010) These MAMPs are recognised by Toll-like receptors, and a play a key role in antigen recognition by the innate immune response (Smith *et al.* 2003). The *Campylobacter* flagellin is also heavily glycosylated, producing a 6kDa shift from the protein's predicted molecular mass. A total of 19 serine/threonine residues – all of which are located on the surface exposed domain – have modified glycosyl residues, with several modifications involving pseudaminic acid derivatives mediated by the *pseA-I* genes (Thibault *et al.* 2001).

1.3.6.2. Chemotaxis

Chemotaxis is the directional movement mediated by the presence of attractive or repulsive gradients. Prokaryotes have a chemosensory pathway referred to as the histidine aspartate phosphorelay (HAP), which is able to modulate the direction in which the flagellar motor spins – and thus generate random directional change – in response to chemotactic stimuli (Wadhams & Armitage 2004).

The *Campylobacter* chemoreceptor system is composed of transducer like proteins (Tlps), a membrane-associated histidine kinase (CheA), a linker protein (CheW) and cytoplasmic response regulator (CheAY) (Yao *et al.* 1997, Lertsethtakarn *et al.* 2011). Genomic analysis has also revealed additional subunits involved in chemotaxis, notably including CheV, which is homologous to CheW (Alexander *et al.* 2010, Marchant *et al.* 2002). *Campylobacters* response regulator CheAY is atypical in that it features a response regulator domain (homologous to the CheY protein) and the C-terminus of CheA subunit (Zautner *et al.* 2012).

C. jejuni has 10 identified transducer-like proteins (Tlp1-10) and two aerotaxis genes (*Aer1, 2*) belonging to 3 groups which are responsible for ligand binding. Group A Tlps (Tlp1-4, 7 and 10) - responsible for the sensing of extracellular ligands - are highly conserved across *C. jejuni* strains and consist of distinct cytoplasmic, periplasmic and transmembranous regions (Zautner *et al.* 2012). Group B contains the singular Tlp9, which is also referred to as CetA for *Campylobacter* energy taxis A; CetA lacks any ligand-binding domain and instead acts as a signal transducer to the aerotaxis proteins Aer1/2. Group C (Tlp5, 6 and 8) are purported to be responsible for cytoplasmic signal detection (Zautner *et al.* 2012, Lertsethtakarn *et al.* 2011, Walker *et al.* 1986, Young *et al.* 2007).

CheW is a scaffolding protein that links Tlps to CheA in the quaternary Tlp-CheW-CheA complex and is involved in signal transduction. Upon formation of the oligomeric quaternary structure, CheA undergoes autophosphorylation (Gegner *et al.* 1992, Korolik *et al.* 2008). CheY then binds to CheA and the phosphoryl group is transferred. CheY phosphorylation permits interaction with FliM – a motor switch protein – and changes the direction in which the flagellar filament is rotated (Korolik *et al.* 2008).

During host colonisation in both commensal and pathogenic contexts, *C. jejuni* initially establishes itself in the mucus layer covering the intestinal epithelium (Young *et al.* 2007). The intestinal mucus layer is formed by the secretion of mucin glycoproteins from epithelial cells. A major component of these glycoproteins is L-fucose, which is located in the terminal position of the oligosaccharide chain (Becker & Lowe 2003, Lindén *et al.* 2008). L-fucose acts as a chemoattractant for *Campylobacter*, and is hypothesised to mediate chemotaxis towards the mucus layer in *C. jejuni* infections (Stahl *et al.* 2012). Additionally, a number of *Campylobacters'* preferred nutrient sources (*i.e.* amino acids and TCA-cycle intermediates) also produce a chemotactic response and are common components of the gut environment. *C. jejuni* attractants include, but are not limited to L-fucose, L-cysteine, L-glutamate,

L-serine, L-aspartate, succinate, fumarate, pyruvate, α-ketoglutarate and malate (Hugdahl *et al.* 1988).

1.3.6.3. Adhesion

Following localisation to the intestinal mucus, *Campylobacter* must adhere to host cells to overcome peristaltic motion and excretion (Baqar *et al.* 2008). Unlike other Gram-negative enteropathogens, adhesion is not mediated by the production of fimbrial structures, and instead relies on a complex network of cellular surface structures known as adhesins (Guerry 2007).

Campylobacter has several suspected adhesins, including lipooligosaccharides, outer-membrane proteins, and flagella, however there are only a handful of examples that have been confirmed via *in vitro* study (Flanagan *et al.* 2009, Golden & Acheson 2002). For example, two genes - *cadF* and *jlpA* – have been shown to specifically bind to host fibronectin receptors, with mutagenesis studies showing reduced adhesion in mutant phenotypes (Golden & Acheson 2002, Flanagan *et al.* 2009). *C. jejuni* adhesion is thought to occur at the apical surface of epithelial cells in the perijunctional area prior to host-cell invasion (Hu & Kopecko 2008). The amount of unconfirmed adhesins is far greater and continues to be a keen topic of research (Ramirez-Hernandez *et al.* 2015, Lugert *et al.* 2015, Boehm *et al.* 2015).

1.3.6.4. Invasion

There is conclusive evidence from human biopsies, animal infection models and intestinal epithelial cell lines that *Campylobacter* invades host cells (Poly & Guerry 1989, Yao *et al.* 1997, Szymanski *et al.* 1995). Several mechanisms of cell entry have been described, and are largely impacted by strain and cell line, the pathway described has been studied in *C. jejuni* 81-176 (Hu & Kopecko 1999).

Campylobacter initiates invasion by secreting effector proteins; a study identified that a minimum of 8 novel *C. jejuni* proteins are synthesised following cellular adhesion; in particular, the *ciaB* gene was identified as

a key component of the secretion pathway, with a total loss of adhesion-dependant protein secretion in knockout mutants (Konkel *et al.* 1999). The secreted proteins function by triggering the depolymerisation of actin filaments, leading to the formation of microtubule based protrusions and ultimately leads to endocytosis (Kopecko *et al.* 2001, Lugert *et al.* 2015).

Internalised *C. jejuni* cells are contained within a membrane bound vacuole (Lugert *et al.* 2015). The mechanism underpinning cellular entry limits the number of endocytosed bacteria to one or two organisms per cell, though limited intra-vacuole replication has been shown to occur (Hu & Kopecko 1999). The vacuole then translocates along intracellular microtubules to the basolateral cell surface using a dynein motor protein. At the basal surface, cells are exocytosed in as little as 15 minutes following invasion and are able to regain entry to the epithelium basolaterally (Hu & Kopecko 2008).

Diarrhoea symptoms are purportedly generated due to a collection of factors. During the adhesion and invasion process, there is an increase in intracellular Ca^{2+} and activated protein kinase C which has been linked to an increase in fluid secretion (Hu *et al.* 2005). Additionally, *Campylobacter* has been shown to secrete a number of proteins (Wassenaar 1997). Amongst the more well defined proteins is a cytolethal distending toxin (CDT) that arrests the cell cycle and leads to intestinal cell death. CDT-mediated cytotoxicity activates inflammatory pathways via the induction of IL-8 secretion, which coupled with the increased fluid secretion and impaired absorption capacity of the damaged intestinal epithelium leads to diarrhoeal stools (O Cróinín & Backert 2012, Nielsen *et al.* 2012).

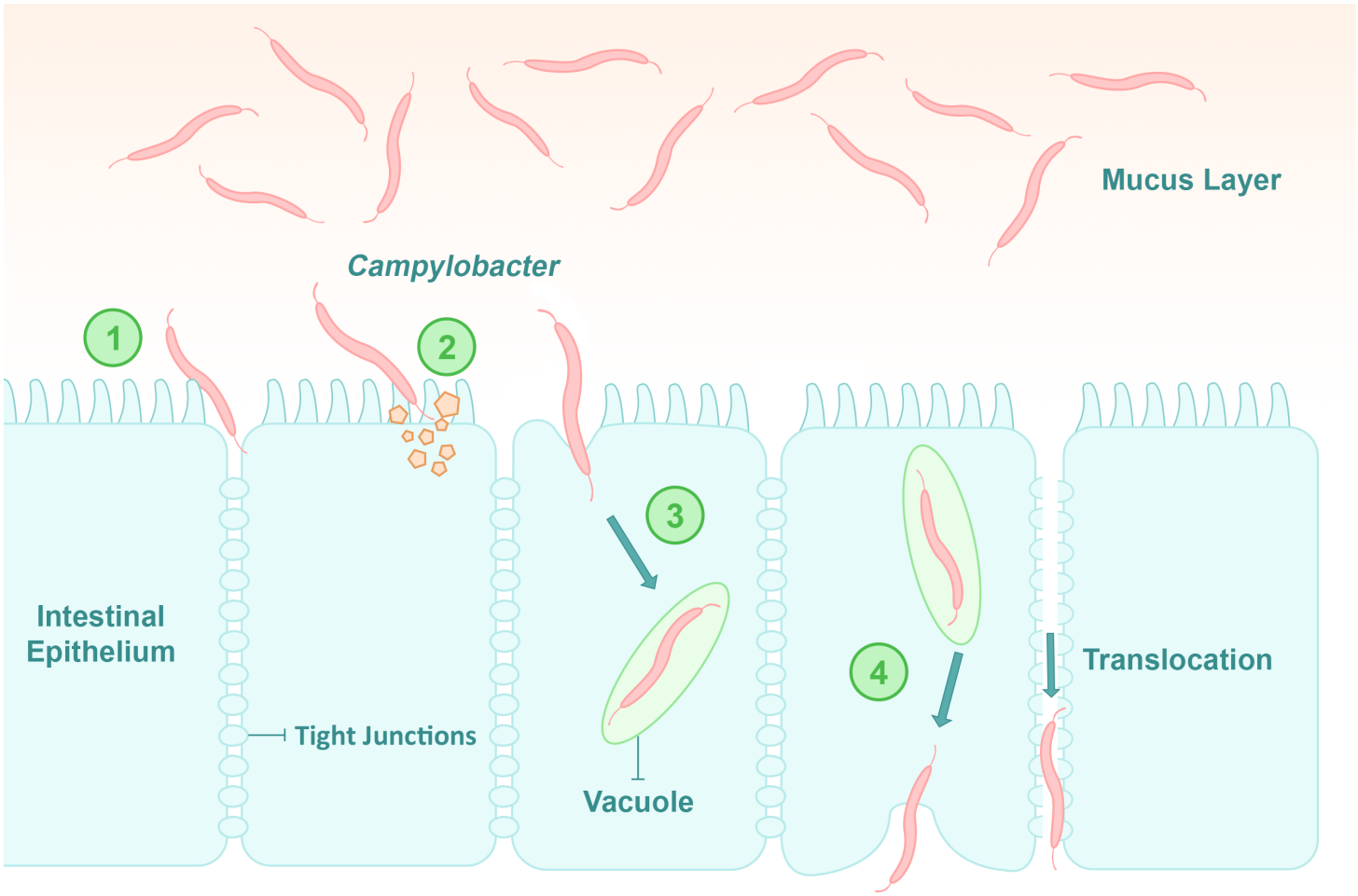

Figure 1.4 - A brief overview of *Campylobacter* pathogenesis

(1) *Campylobacter* cells present in the mucus layer adhere to intestinal epithelium perijunctional regions, mediated by surface bound adhesins (*e.g.* CadF). (2) Bacteria secrete proteins that propagate cellular invasion (*e.g.* CiaB). (3) In response to the secreted invasion effectors, a cascade is triggered that results in the endocytosis of *Campylobacter* cells, which are engulfed in a vacuole. (4) The bacterium moves along microtubule filaments to the basal surface where it is exocytosed. *C. jejuni* has also been shown to disrupt the tight junctions linking the intestinal epithelium. It is hypothesised that *Campylobacter* may exhibit paracellular invasion via translocation of the disrupted junctions. Figure design based on (Kopecko *et al.* 2001).

1.3.7. Genomics

The first complete *Campylobacter* genome sequence of *C. jejuni* clinical isolate NCTC 11168 was published in 2000, using a combination of dye-terminator sequencing and PCR verification (J Parkhill *et al.* 2000). This genome was re-annotated in 2007 to include up-to-date references and methodologies (Gundogdu *et al.* 2007). In the past decade several additional complete genomes have been sequenced, including *C. jejuni* isolates NCTC 11828, 81-176, RM1221, and CG8486 (Hofreuter *et al.* 2006, Pearson *et al.* 2007, Poly *et al.* 2007b, Fouts *et al.* 2005).

The genome of NCTC 11168 is 1,641,481 bp in length; it has a relatively low G+C content of 30.6%, and is configured in a single circular chromosome. *Campylobacter* has a particularly dense genome, with 94.3% of nucleotides distributed across 1643 CDSs. *C. jejuni's* genome is unique in that it does not contain insertion sequence elements, or transposons (J Parkhill *et al.* 2000).

The complete genome sequence of another *Campylobacter jejuni* isolate - RM1221 - was published in 2005 (Fouts *et al.* 2005) and showed a number of differences to the genome of 11168. RM1221's genome is notably longer, at 1,777,831 bp in length and has a total of 1,884 predicted CDSs. The major difference is that RM1221 has four integrative elements, including an inducible *Campylobacter* Mu-like phage (CMLP) and *C. jejuni* integrated elements 1-3 (CJIE) (Clark & Ng 2008). With the exception of CJIE3, several CDSs within these genomic islands are putatively identified as phage-related. Bacteriophages are vectors for horizontal gene movement which can confer an increase in fitness and virulence (Hendrix *et al.* 2000).

During the sequencing of *C. jejuni* isolate 81-176 - the same isolate used in the human trials by Black in 1981 – two plasmids were identified: pVir and pTet (Hofreuter *et al.* 2006). Between 19-53% of *C. jejuni* strains have

been shown to contain plasmids, with many being exclusively transmissible amongst *Campylobacter spp*. (Bacon *et al*. 2000). The pVir plasmid in the 81-176 genome is 37,468 bp in length with 35 open reading frames. Several genes within the plasmid are homologous to type IV secretion system proteins. The plasmid differs from typical *C. jejuni* DNA in that only 83% of the DNA spans coding information, and has a lower GC content at 26%. pVir has been significantly associated with the presence of blood in patient stools, which may reflect an increased level of invasion (Tracz *et al*. 2005).

The *Campylobacter* genome contains several instances of hypervariable mononucleic polyG/C tracts. Ranging from 8-13 bp, these tracts mediate on/off expression of phase-variable genes (known as contingency genes) via slipped-strand mispairing, with a similar system observed in *Neisseria* (J. Parkhill *et al*. 2000). A number of genes controlled in this manner are located in regions responsible for surface structure and likely reflect the mechanism that underpins *Campylobacter*'s diverse surface configurations.

1.3.7.1. MLST

Multilocus sequence typing (MLST) is a genotypic method that unambiguously groups genetically homologous isolates based on the nucleotide sequences of several highly conserved 'housekeeping' genes (Maiden *et al*. 1998).

The MLST scheme for *Campylobacter* was established in 2001 and utilises seven loci: *aspA, glnA, gltA, glyA, pgm, tkt* and *uncA* (Dingle *et al*. 2001). This scheme assigns each loci with an arbitrary allele number based on iterative differences in the gene sequence. The combination of allele numbers forms a unique allele profile that is assigned to a sequence type (ST). Sequence types can be grouped into larger, epidemiologically related lineages known as clonal complexes (CC) that are derived from a single common ancestor (Maiden *et al*. 1998).

For a novel sequence type to be assigned to an existing *Campylobacter* CC it must share a common allele profile at least 4 of the 7 loci. Each CC is labelled with the founding sequence type identifier, for example CC ST-21 originated with a single ST-21 isolate that in itself was assigned a novel sequence type due to differences in 4 or more loci when compared to other isolates within the database (Jolley & Maiden 2010, Dingle *et al.* 2001).

1.3.7.2. Molecular Epidemiology

The introduction of MLST allowed characterisation of *Campylobacter* at the population level, and has revealed a number of attributed sources of clinical infection. Several studies have shown members of the ST-21 complex as the most frequently isolated lineage in clinical samples across multiple countries (Sheppard *et al.* 2009, Kovanen *et al.* 2014, Cody *et al.* 2012, Schielke *et al.* 2014). The ST-21 lineage has been described as a supercomplex due to an abundance of constituent sequence types (Dingle *et al.* 2001). At the time of analysis, 684 unique sequence types were assigned to the ST-21 complex, which accounts for 17.4% (7858) of all isolates submitted to the PubMLST database (Jolley & Maiden 2010).

Members of the ST-21 lineage – and other common clinical sequence types such as ST-45 and *C. coli* ST-828 - have been isolated from a variety of hosts. Rapid transmission events between these reservoirs (and the large number of sequence types within some complexes) erodes signals of host-association (Dearlove *et al.* 2015), however the frequency at which certain complexes are isolated from clinical samples suggests that some isolates have niche adaptations that favour the development of human pathogenesis (Olson *et al.* 2008). Figure 1.5 shows the distribution of clinical cases attributed to the major CC's submitted to BigsDB as of October, 2016.

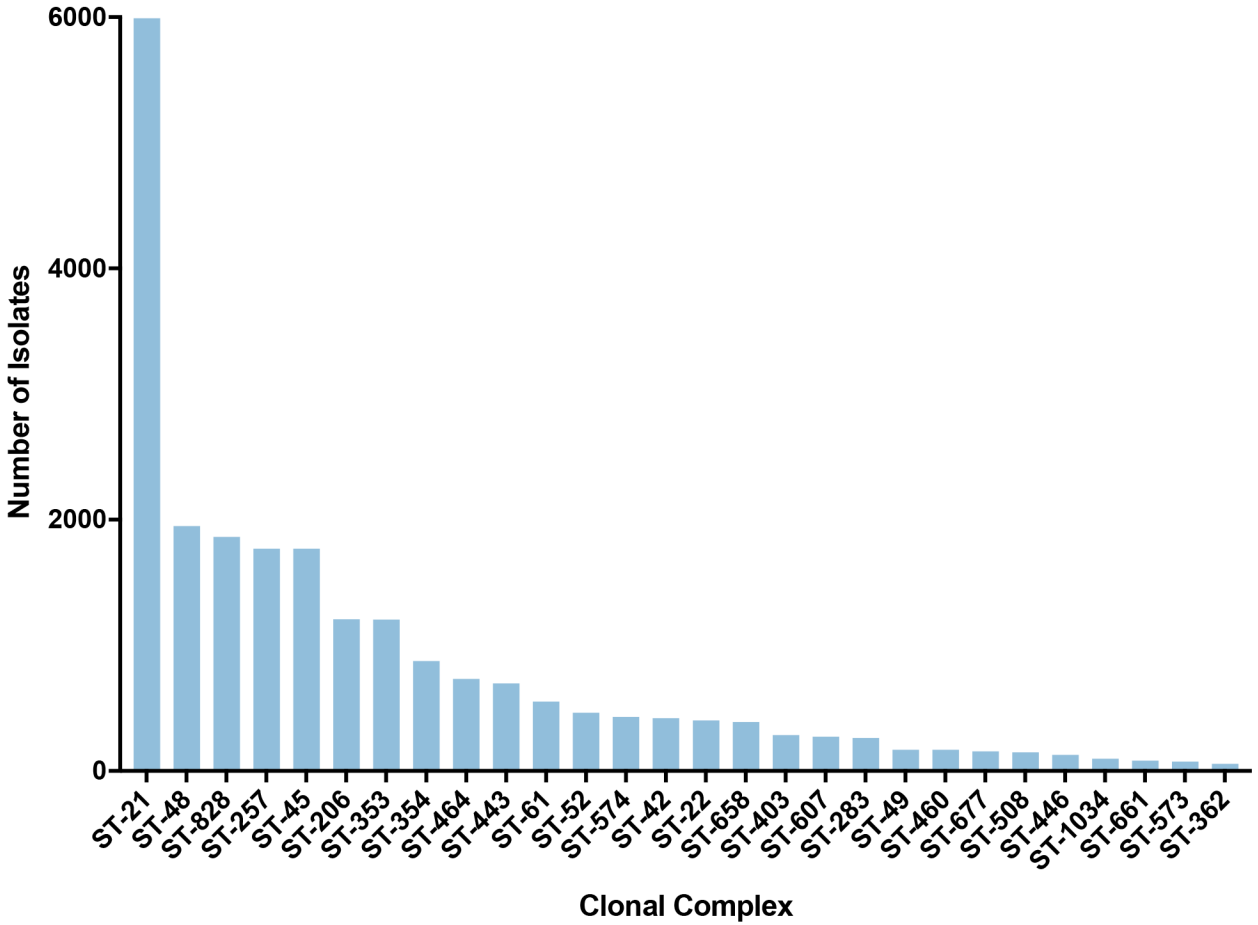

Figure 1.5 - Total number of isolates from identified clonal complexes submitted to BigsDB isolated from human stool

Figure produced from public data listed on the PubMLST database for *C. jejuni* and *C. coli*. Inclusive of clonal complexes with more than 50 isolates, matching sources: human stool, blood culture, CSF and unspecified (Accessed 12/06/17).

1.3.7.3. Genetic Diversity

Genetic diversity, and by extension bacterial evolution as a whole, is dominated by two key factors: the rate of mutations that arise as a result of DNA damage or replication error, and the acquisition of external variation via horizontal or lateral gene transfer. The former produces a gradual clonal divergence, with diversity accumulating over a long period of time, whilst the latter can import a dramatically different stretch of DNA in a single genetic event, significantly altering any single clone (Sheppard & Maiden 2015).

C. jejuni displays a dramatic genotypic diversity which is largely thought to be due to its propensity for both inter and intraspecies recombination. *Campylobacter* is naturally competent, and is able to import and incorporate environmental DNA into its genome. More than 40 bacterial species are naturally competent, including *Neisseria gonorrhoeae* and *Helicobacter pylori* (Sparling 1966, Tsuda *et al.* 1993). *Campylobacter's* ability to share genomic DNA is well documented (Iraola *et al.* 2014, Wells & Bennik 2003, Wassenaar *et al.* 1995), As a result, the population structure is weakly clonal, and generally grouped into larger epidemiologically related lineages.

Sheppard *et al.* (2008), demonstrated that *Campylobacter* has undergone a recent increase in bidirectional recombination between *C. jejuni* and *C. coli,* which may be resulting in the convergence of the two species. This is particularly apparent in *C. coli,* which undergoes asymmetric gene flow and was demonstrated to import *C. jejuni* alleles across all identified loci (Sheppard *et al.* 2008). This increase is likely as a result of agricultural activities, which have removed ecological barriers and allowed the two species to interact. This highlights *Campylobacters* ability to rapidly evolve as a result of its proclivity for recombination.

The amount of diversity observed in *Campylobacter spp.* may mask its epidemiology. Studies have identified the presence of genetic diversity through repeated sampling of a single infected person (Cody *et al.* 2013).

All isolates were of the same sequence type, and therefore assumed to be the same clone of *Campylobacter* responsible for initiating the infection. In some patients, variants in as many as 14 loci were detected. The same study aimed to identify instances of related cases amongst a dataset of 379 clinical isolates. The researchers assigned the cut off value for describing a case as linked at 20 loci.

As this study is based on wgMLST, whereby each loci is allocated an allele number by the same mechanism as ST assignation, the measure of variable loci is only quantitative. A single polymorphism in any loci would be sufficient to generate a different allele and thus categorise a locus as variable, which is indistinguishable from a loci containing multiple polymorphisms. It is therefore unclear as to how this methodology would compare to other measures of diversity such as single nucleotide polymorphism (SNP) typing. For example, if an isolate contains X number of called variable loci (X<20) and Y number of polymorphisms, the isolate may be considered as related via SNP typing where X=Y, or drastically different if Y is sufficiently greater than X. Equally, isolates with 21 SNPs in unique loci would be sufficient to discount linkage via wgMLST methods. Given the vast diversity observed amongst clonal isolates derived from a single patient, and the propensity of *Campylobacter* to transfer genes, it is not unreasonable to assume that case linkage may occur and is simply
not detected.

1.3.8. Study Rationale

The majority of *Campylobacter* infections are thought to be sporadic in nature, and epidemiological studies have not definitively identified any instances in which case linkage has been observed. However, these studies have relied on comparative techniques that exclusively rely on *de novo* genomic assembles.

As a product of the *de novo* assembly process, the resulting genome is a consensus estimate sequence that has been collapsed to a depth of 1x. Microbial genomic reads typically have a coverage depth of

approximately 30x – that is to say that any given position in the genome has been sequenced 30 times. By comparing these read files to a target *de novo* assembly (*i.e.* read mapping), a more thorough and definitive measure of variation between an isolate pair can be obtained.

Read mapping approaches produce a list of genomic positions at which a non-congruent base call occurs. A variation at a given base position is known as a single nucleotide polymorphism (SNP). At a superficial level, the quantity of polymorphisms observed between two isolates can be used as an indication of homology, with fewer polymorphisms indicating a greater degree of similarity. The individual positions of the polymorphisms and their distribution across the genome can also be used to delineate case linkage, which is currently not possible with wgMLST. For example, by the wgMLST scheme, an isolate pair may contain several hundred polymorphisms across 19 loci and be considered to be linked. By the same methodology, an isolate pair could contain just 21 polymorphisms in unique loci and be assigned as non-linked.

This study aims to utilise a read mapping approach to identify any potential linkage from a set of human clinical campylobacteriosis cases, which may otherwise be missed via conventional (*i.e.* wgMLST) genomic methods. Utilising a read mapping approach will provide a thorough and objective methodology for comparing genomic sequences, and definitively identify any potential instances of case linkage. It will also be sensitive to variation at intergenic positions, which is not possible via wgMLST analysis.

Additionally, at present the abundance of CC's is thought to be equally representative of the United Kingdom as a whole. As there have been no genetic epidemiology studies of *Campylobacter* in Nottingham, the wider population structure of clinical *Campylobacter* across Nottinghamshire will be elucidated using phylogenetic analyses. This will provide an insight into the abundancies of clinically associated complexes and allow comparison between genomic studies elsewhere.

Chapter Two: Methods

2.1. Bacterial Strains

All bacterial isolates were obtained from Nottingham University Hospital's Queen's Medical Centre (QMC) - a large health trust that services the greater Nottinghamshire area. Their pathology lab processes approximately *800 Campylobacter spp*. isolates annually, and for the purposes of this study have captured both the original stool sample and a frozen stock culture of the organisms.

All analysed strains were obtained from cryopreserved samples stored at -80°C, which were prepared directly from an initial stool culture on a mCCDA based selective agar. All isolates were transported on ice from QMC to the CL2 laboratory at Clifton Campus, Nottingham Trent University in less than 30 minutes, and again stored at -80°C until use.

Both the initial cryopreserved samples and sub-stocks were stored using Microbank™ Preservation Beads (Prolab Diagnostics, United Kingdom, PL.170). These preservation beads are chemically coated to promote bacterial adhesion, and are suspended in a cyropreservative that is optimised for fastidious organisms.

2.2. Sampling

Queen's Medical Centre provided monthly reports on all *Campylobacter* isolates collected by their microbiology department. These reports included a sample identifier, the date of sample collection, and the date on which the sample was received.

The sampling period covered a 365-day range from January 1st – December 31st 2014, and included all isolates with a collection date within that time frame. An approximate, raw number of isolates to be sampled each month was determined using local historic incidence data supplied by QMC from recent previous years (provided in Appendix 9.2) using the equations listed on the following page.

$$\frac{Average\ Yearly\ Incidence}{Total\ Sequencing\ Capacity} = x$$

$$\frac{Average\ Monthly\ Incidence}{X} = Raw\ Sample\ Number$$

The average daily incidence observed in 2014 was compared to the historic data to obtain a suitable sample size that accounted for any deviation or natural fluctuation in isolation numbers, and allow the sequencing capacity to be distributed appropriately. A daily average was used over monthly average to account for and normalise leap years.

$$\sum \frac{Daily\ Reported\ Isolates}{No.of\ Days\ in\ Month} = Daily\ Average$$

$$\frac{2014\ Daily\ Average}{Historic\ Daily\ Average} = Sampling\ Multiplier$$

$$Raw\ Sample\ Number \times Sampling\ Multiplier = Corrected\ Sample\ Number$$

2.3. Culture Media

Isolates were cultured directly onto *Campylobacter* Blood-Free Selective Agar Base (Oxoid, United Kingdom, CM0739) – a modified Charcoal-Cefoperazone-Deoxycholate agar - supplemented with 16mg/500ml of Cefoperazone and 5mg/500ml of Amphotericin B (Oxoid, United Kingdom, SR0115).

The agar was prepared by suspending 22.75 g of the agar base in 500 ml of distilled water and mixed using a magnetic stirrer. The solution was sterilised by autoclaving at 121°C for 15 minutes, and stored at 55°C until required. Before pouring, the agar was again magnetically stirred whilst cooling.

The selective supplement was prepared by reconstituting the powdered antibiotics (stored at 4°c) in 2 ml of sterile distilled water and added to the cooled agar before a final round of magnetic stirring. Plates were poured immediately under aseptic conditions and used within 48 hours.

2.3.1. Culture Conditions

The cryopreserved bacterial cultures were removed from the -80°C freezer and thawed on ice for approximately 10 minutes before use. Care was taken to ensure a minimal amount of freeze thawing occurred, and that samples were not thawed for longer than necessary.

A single Microbank preservation bead was transferred from the storage tube onto pre-prepared mCCDA plates. A 10 µl inoculation loop (Scientific Laboratory Supplies, United Kingdom, SLS2010) was used to smear the bead across the surface of the plate in a simple streak pattern.

Isolates were incubated for approximately 48 hours at 37°C in an anaerobic workstation (Don Whitley Scientific, United Kingdom, DG250), configured with a microaerobic gas canister (5% O_2, 10% CO_2 and 85% N_2, BOC, United Kingdom).

Samples that failed to produce sufficient growth were re-cultured in Mueller Hinton Broth (Oxoid, United Kingdom, CM0405), which was supplemented with the previously described selective formula (Oxoid, United Kingdom, SR0115).

The broth was prepared by mixing 10.5 g of the broth base with 500 ml of distilled water and sterilised by autoclaving at 121°C for 15 minutes. The broth was allowed to cool before adding the reconstituted selective supplement and subsequently aliquoted into sterile 15 ml conical tubes (Sarstedt, Germany, 62.554.002).

A single Microbank preservation bead was transferred directly into a conical tube containing 10 ml of MHB. The tubes were sealed and transferred to a shaking platform inside the anaerobic workstation; once all tubes were seated correctly, the conical tube lids were loosened to allow sufficient gas permeability and the shaking platform was set to 100 rpm. Isolates were incubated for 24 hours and then sub-cultured onto mCCDA for a further 24 hours as previously described.

Samples that continued to yield insufficient growth following pre-enrichment were discarded and an alternative sample was used. Figure 2.1. gives a summary of the isolation process.

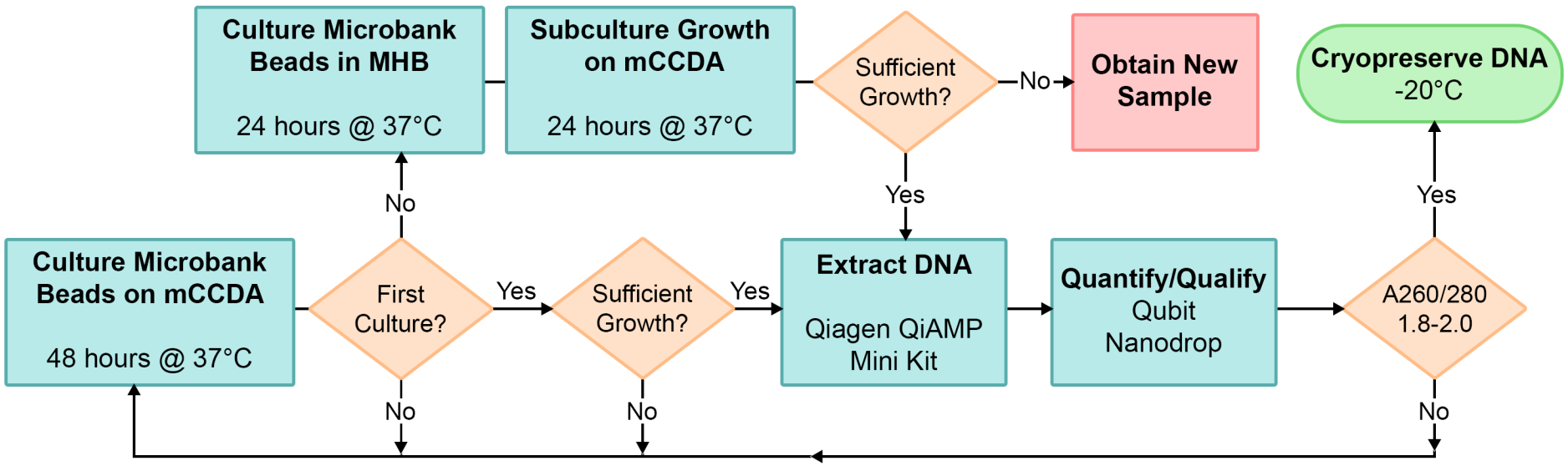

Figure 2.1- Summary of sample preparation and DNA qualification process.

Cultures that failed to produce sufficient growth during the first round of culture were preenriched in MHB. Isolates that continued to provide insufficient growth were discarded. DNA was qualified based on the absorption ratio of the bulk sample, with an A260/280 ratio of 1.8-2.0 treated as acceptable for WGS.

2.4. DNA Isolation

DNA was prepared for whole genome sequencing (WGS) using the QIAmp DNA Mini Kit (Qiagen, Netherlands, 51306). The standard assay protocol provided in the QIAmp DNA Mini and Blood Mini Handbook (Qiagen, Netherlands, HB-0329-004) - summarised in Figure 2.2– was modified to account for *Campylobacter* biology.

Due to the presence of numerous endonucleases in *Campylobacter,* it was important to ensure protein denaturation occurred quickly after cell lysis to prevent nucleic acid degradation. Initial cell lysis with 'Buffer AW' was limited to 30 minutes, followed by a 60-minute incubation with proteinase K (20 mg/ml, 600 mAU/ml).

During the proteinase K incubation stage, the sample was frequently vortex mixed to ensure efficient enzyme-substrate binding. Vortexing also helped to disassociate tightly bound pellets that formed as a result of autoagglutination. At the end of the 60-minute incubation period samples were checked for homogeneity. Any samples that were not fully digested (*i.e.* if a visible mass was still present) were returned for incubation and reassessed in 10 minute intervals.

The digested solution was treated with 20 µl of 20 mg/ml RNAse A solution (Sigma Aldrich, United Kingdom, R6148) and incubated at room temperature for 15 minutes.

To avoid mechanical shearing, the DNA extraction was transferred to the silica columns by gently pouring the bulk of the tube contents. Remaining excess solution was pipetted across using a P1000 tip.

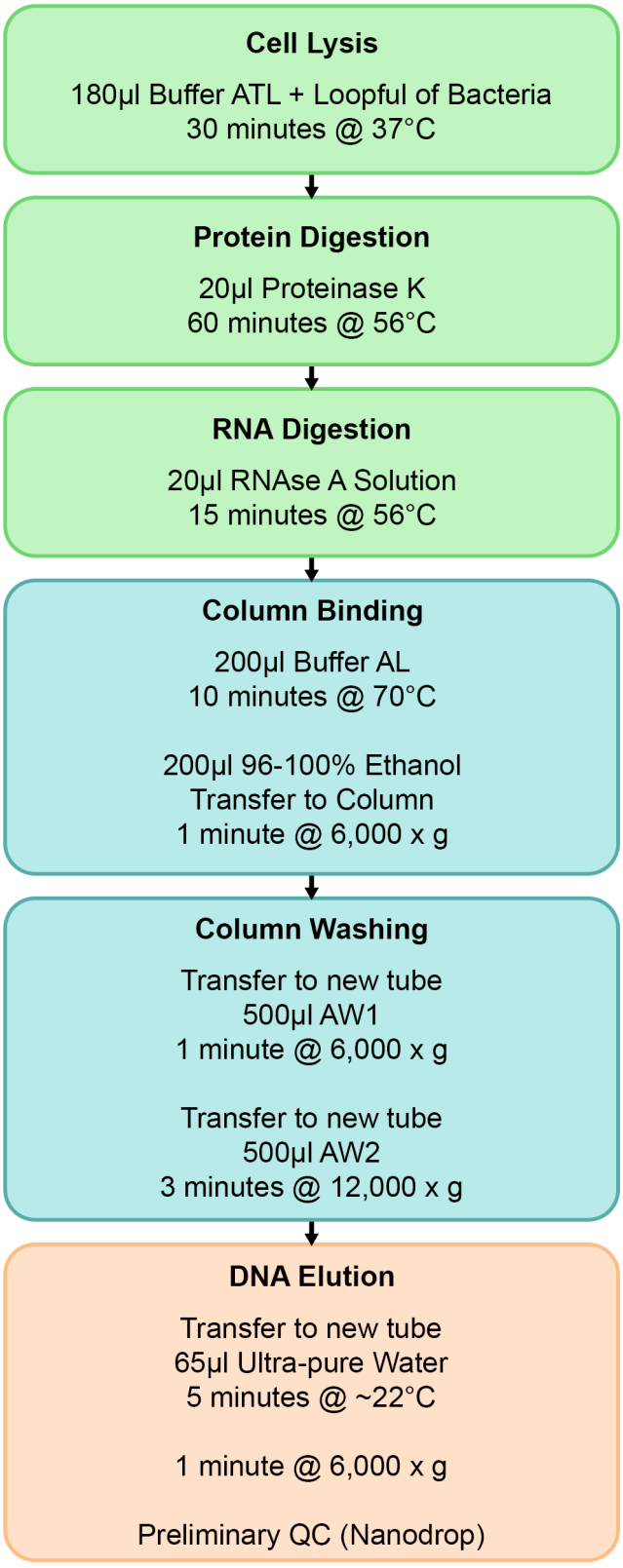

Figure 2.2 - Qiagen QIAmp DNA Mini Kit DNA Isolation Workflow

Green boxes contain processes related to bacterial cell lysis and digestion, blue relate to processes wherein DNA is bound to the silica columns and orange to DNA elution and handling.

As the elution buffer provided in the kit (Buffer AE) contains EDTA – an inhibitor of downstream sequencing applications – ultrapure sterile distilled water was used in its place. DNA was eluted in a reduced volume (65 µl vs. recommended 200 µl) to increase stock concentration, and allow re-elution of DNA from the column if any contamination was subsequently observed. DNA was stored in sterile 1.5ml Eppendorf tubes at -20°C.

2.4.1. DNA Qualification

The quality of DNA extractions was initially verified using the Nanodrop 2000 spectrophotometric platform (ThermoFisher Scientific, USA, ND-2000) to identify samples with sufficient yield, quality and purity. The Nanodrop assesses quality based on the absorption spectra of nucleic acid, with acceptable quality thresholds inferred from absorption ratios. DNA extractions that had a 260:280 nm ratio of between 1.80-2.00 and a 260:230nm ratio of between 2.00 and 2.20 were deemed to be sufficiently pure for WGS.

Any samples with abnormal ratios or a non-uniform curve were diagnosed for likely contaminants. Where suitable, the elution stage of the DNA extraction step was repeated to yield a purer preparation. Extractions that did not meet the quality threshold after troubleshooting were discarded and repeated.

The Nanodrop platform also quantitates DNA. As the quantification is non-specific for DNA, and instead reports on all substances that absorb at 260 nm (*e.g.* RNA), the quantification was only used to ensure preparations had sufficient approximate yield. DNA preparations in excess of 30 ng/µl were treated as appropriately concentrated.

2.4.2. DNA Quantification

To obtain an accurate quantification of double-stranded DNA (*i.e.* sequencable material), samples were measured using the Qubit fluorometric platform (Thermofisher Scientific, USA, Q33217). This assay uses a dye that specifically binds to ds-DNA, forming a complex that is highly fluorescent.

For initial stock measurements, 2 µl of DNA was measured using the broad-range assay kit (Thermofisher Scientific, USA, Q32850) rated at a range of 2-1000 ng (*i.e.* measurable concentration range 1-500 ng/µl).

Subsequent measurements of dilutions at low concentrations utilised the high sensitivity assay kit (Thermofisher Scientific, USA, Q32851). This kit is rated at a range of 0.2-100 ng, and is less accurate at the extremes of that range. When diluting to the required input concentration for WGS (0.2 ng/µl), the amount of DNA was increased to 10 µl to increase the accuracy and sensitivity of the assay.

2.5. Whole Genome Sequencing

Whole genome sequencing was conducted on the MiSeq platform (Illumina, USA) in two tranches, one of which was conducted externally by Ben Bascoe and Samuel Sheppard at Swansea University. The methodology described herein is largely representative of both tranches, though minor variations may exist.

Prior to library generation, a subset of DNA samples was verified using gel electrophoresis to ensure no degradation had occurred during freeze storage. Samples were treated as intact if the electrophoresis produced high molecular weight, non-smeared bands. The DNA was then transferred into a 96-well plate (Corning, Netherlands, AXP470) and diluted to 0.2 ng/µl using ultrapure sterile distilled water.

Genomic libraries were prepared using Illumina's NextEra XT kit (Illumina, USA, FC-131-1024) according to the protocol described in the Nextera XT DNA Library Prep Reference Guide (Illumina, USA, 15031942 v01). Briefly, the dilute DNA samples were incubated with a dual tn5 transposase mix that fragments dsDNA and adds an adaptor sequence required for PCR amplification. Libraries were indexed using Illumina's Nextera XT Index Kit V2 (Illumina, USA, FC-131-2001), with resulting PCR products cleaned using Agencourt AMPure XP beads (Beckman Coulter, USA, A63880).

In lieu of the bead based normalisation method provided with the library preparation kit, libraries were normalised manually. The bead normalisation protocol requires a high concentration of DNA, and uses the total amplification product. Manual normalisation allowed a finer control over the final library concentration and produced a greater number of successful library preparations, as well as providing excess product for QC and contingency purposes.

To manually normalise the libraries, concentration of the cleaned PCR product was determined using the Qubit HS Assay Kit as previously described. Average library insert size was determined by running a subset of samples on an Agilent 2200TapeStation (Agilent Technologies, USA, G2964AA) using a D1000 high sensitivity ScreenTape and reagents (Agilent Technologies ,USA, 5067-5582 + 5067-5583). The average insert size provided by the TapeStation was used to inform the molar conversion (*i.e.* ng/µl to nM) using the following equation:

$$\frac{Concentration\ in\ ng/\mu l}{(660\ \mathrm{g/mol} \times Average\ Insert\ Size)} \times 10^6$$

The libraries were diluted to 4nM and denatured using sodium hydroxide as described in the MiSeq System Denature and Dilute Libraries Guide (Illumina, USA, 15039740 v01). Due to the required depth of coverage for 96 *Campylobacter* genomes, libraries were loaded onto a MiSeq Reagent Cartridge v3-600 (Illumina, USA, MS-102-3003) at a final concentration of 20 pM. The MiSeq was configured to run paired end 300 bp reads in 'FastQ generation' mode.

2.5.1. Estimating Sequencing Coverage

Expected sequencing coverage can be estimated using the equation given below. The MiSeq's V3 chemistry specification gives an approximate coverage of 47.98 vs. 23.99 on the V2 chemistry. Multiple factors such as the number of clusters passing filter and resulting per base quality will impact the final usable coverage; coverage is also not distributed equally across the length of the genome, and as such the V3 chemistry was selected to err on the side of caution.

$$\frac{Read\ Length \times Number\ of\ Reads}{Genome\ Length \times Number\ of\ Libraries} = Coverage$$

$$MiSeq\ V2\ Chemistry\text{: } \frac{250 \times 15,000,000}{1,628,115 \times 96} = 23.99$$

$$MiSeq\ V3\ Chemistry\text{: } \frac{300 \times 25,000,000}{1,628,115 \times 96} = 47.98$$

2.5.2. Genomic Read Curation

Following FastQ generation, the run performance was analysed using Sequence Analysis Viewer (Illumina, USA) to provide an overview of the general read quality. Aspects such as the overall and per-cycle % of bases >Q30, the percentage and distribution of reads mapped to an index ID, and channel intensity were used to judge the run quality.

FastQC was used to produce individual sample reports (Andrews, 2010). Samples that exhibited a drastically declining read quality were processed using Sickle, a paired-end, sliding window quality trimmer (Joshi and Fass, 2011). Samples that showed residual adapter sequences – that are typically removed during the FastQ generation process – were processed using Trimmomatic (Bolger et. al., 2014).

2.5.3. Genome Assembly and Annotation

Reads were assembled using SPAdes 3.5.0 (Nurk, Bankevich *et al.*, 2013) due to its computational efficiency and strong performance with small genomes. A script (AutoSPAdes Appendix 9.3) was developed to automatically invoke SPAdes on large FastQ datasets, and produce a collection of final assemblies named according to the sample ID. SPAdes was configured to utilise the 'careful' flag, which runs a mismatch corrector to reduce errors and short indels. Draft assemblies were subsequently processed using QUAST, which generates a report on assembly statistics such as number of contigs and N50 that were used to assess the quality of the assembly (Gurevich *et al.,* 2013).

Annotation was performed using Prokka (Seeman, 2014) under default settings. A script (AutoProkka, Appendix 9.4) was developed to automatically invoke Prokka on a directory of assemblies. The script collects and renames the gff annotation files from each sample into a user specified directory and greatly reduces the hands-on time usually required to annotate large datasets.

2.5.4. MLST Profiling

Draft assemblies were initially analysed using 'mlst' to gain MLST allele numbers and sequence types (Seeman, 2014). Isolates were also referenced against the PubMLST *Campylobacter* database (Jolley and Maiden, 2010), to provide information on Clonal Complex. At the time of analysis there were 8389 MLST profiles defined in the database.

Sequences from isolates belonging to the same complex and sequence type were then aligned to identify SNPs as an indication of genetic homology. Alignment was performed using SMALT 0.7.5 (Welcome Trust Sanger Institute), and the subsequent files were processed using the SAMTools suite of utilities (Li *et al.*, 2009). Pairwise distance matrices were produced using MEGA 6 (Tamura *et al.*, 2013) to quantify sequential distance as a raw number of polymorphisms. For highly similar isolates, allele profiles for all identified loci were scored using all available loci deposited to BigsDB and compared in a pairwise manner.

2.5.5. Core Genome Phylogeny

Phylogeny was reconstructed using the Harvest software suite (Treangen *et al.*, 2014) using NCTC11168 (ST-21) as a reference genome. Parsnp – part of the Harvest suite – performs a core genome alignment based on MUMer identity scores. Parsnp was configured to treat genomes as 'curated' to force inclusion of as many isolates as possible.

Maximum-likelihood trees were produced from the intra-lineage alignments using RAxML (Stamatakis, 2014). The resulting phylogenetic data was visualised using FigTree (Rambaut, 2007) and manually annotated to identify CC types of the major clades.

2.5.6. Deep Sequencing

Several isolates from the clinical dataset were also sent for Deep Sequencing (*i.e.* sequencing to a much higher level of coverage) and analysed using a workflow described by Lieberman *et al.* (2014). The marked increase in coverage provided far greater resolution at a read-by-read basis, and allowed distinction of minor allelic variation that would otherwise be undetectable by conventional SNP typing.

Greater detail is given in chapters 3 and 4, however briefly; FastQ files containing raw sequencing reads were initially filtered using Sickle (Joshi and Fass, 2011). This discarded any reads below 50bp in length, and with a quality score of <20. The filtered reads were subsequently processed with Bowtie2 (Langmead and Salzberg, 2012), which aligned reads to a reference; in this instance, deep-sequenced isolates were mapped against assemblies of their own genome. Bowtie was configured to discard any fragment length (inclusive of gaps) >2000bp in length and any reads in which the frequency of ambiguous characters exceeded 1%. The data was then parsed through the Samtools software suite (Li *et al.*, 2009) removing reads with aberrantly high depth and any alignments that did not have a Q-score of > 30. Due to the inaccuracies of Illumina's SBS technology, insertion and deletion events were discarded.

Samtools' mpileup package produces several statistical metrics by which each called allele was further assessed. More than 3% of reads had to support a minor allele, with at least 15 reads mapping the variant (defined by the AF1 value and DP4 value respectively). Strand bias and tail distance bias were also considered, discarding called alleles with P-values of $< 10^{-5}$. This process filtered any false positive polymorphisms or strand asymmetry introduced by misalignments in the computational pipeline. Figure 2.3 provides an overview of the deepsequencing pipeline.

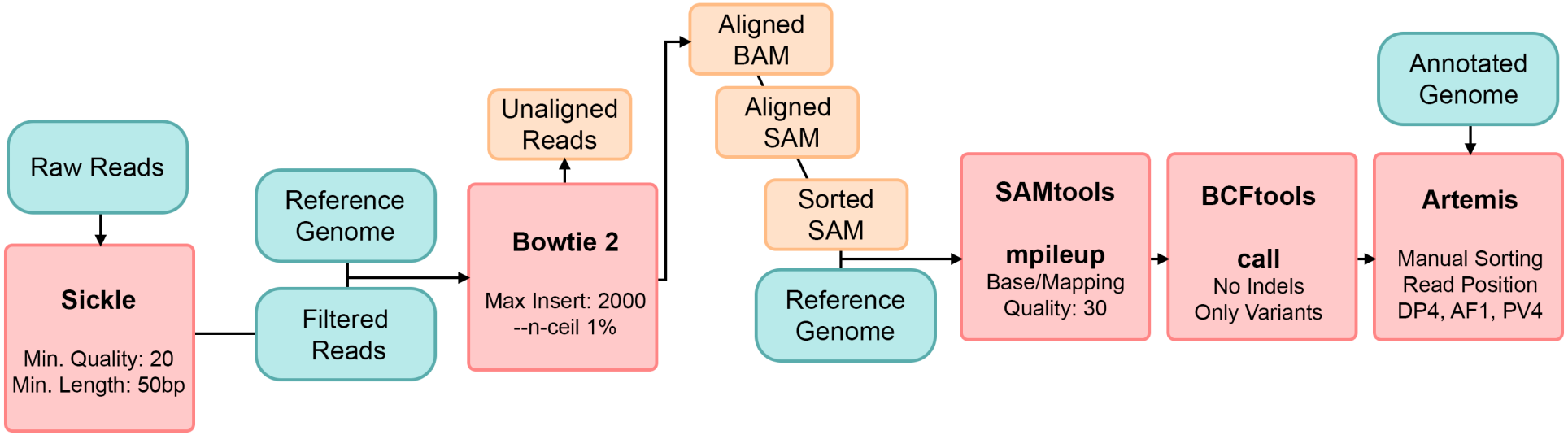

Figure 2.3 - Overview of the deep sequencing analysis pipeline

Software names highlighted within red boxes in bold, with key optional changes to default parameters listed below.

Chapter Three: Genetic epidemiology of clinical *Campylobacter spp.*

3.1. Abstract

Current molecular studies of clinical *Campylobacter* epidemiology are limited to a small number of sites, with population diversity assumed to be representative of that across the United Kingdom. Additionally, links between sporadic cases are yet to be established, with the majority of infections thought to be generated from genetically distinct isolates. However, the techniques employed to study case linkage have thus far relied on incomplete molecular methods (*e.g.* wgMLST) and as such may miss fine scale associations.

Using a higher resolution of analysis at the whole genome level, 158 clinical isolates from the greater Nottinghamshire area were analysed to assess the local population structure and investigate potential case linkages. The distribution of major dominant CC's identified amongst the Nottinghamshire dataset was in agreement with other sites in the UK, verifying the extrapolations from previous studies. Three isolates yielded novel MLST allele profiles, and displayed a core-genome homology to poultry associated ST-573 isolates.

Four instances (2.5%) of case linkage were observed across the dataset when analysed using a read mapping approach. This is the first use of read-mapping as a measure of *Campylobacter* genetic epidemiology. When assessed using previously employed methods alone, several additional instances of case linkage are inferred. However, read mapping reveals a large level of diversity within these isolates that suggests they are not genetically homologous.

During the course of the study, an unusually high incidence was observed during February, 2014. The majority of cases during this month occurred in two week-long tranches that followed Chinese New Year and Valentine's Day. Additional isolates were collected from the infections following Valentine's day, revealing two instances of case linkage. This represented 12% of sequenced tranche isolates, implying that case

linkage may be enhanced following social events. An increase in *Campylobacter* isolation rates during February is also demonstrable in nationally reported data across multiple years, perhaps suggesting an additional recurring peak in incidence that has not been previously described.

In conclusion, though rare, case linkage is observable amongst isolates that are otherwise considered sporadic in nature. Whilst wgMLST is a useful method of analysis when time or expertise are limited, it is not sensitive enough to truly delineate isolate linkage. Rather, a combination of wgMLST as a guide for investigation, followed by iterative read mapping of related isolates provides a more robust measure of linkage by quantifying variation at the sequence level.

3.2. Introduction

More than a decade ago, Roche released the first commercial NGS platform – the 454 Genome Sequencer (Margulies *et al.* 2006). Whilst crude by today's standards, the 20Mb output provided laboratories with a previously impossible access to genome sequencing. In the years following the 454's release, several companies introduced their own systems based on competing chemistries, with Illumina/Solexa's 'Sequencing-By-Synthesis' (SBS) releasing in 2006 and Applied Biosystems/Life Technologies' 'Sequencing-by-Oligo-Ligation-Detection' (SOLiD) in 2007 (Valouev *et al.* 2008, Voelkerding *et al.* 2009).

Whilst initially an expensive venture, several improvements to the chemistry, hardware and software of sequencing platforms as well as advances in library generation have greatly reduced the per-base cost of WGS. For example, Illumina's HiSeq X Ten platform is capable of producing a human genome at a relative cost of $1000, which represents a 10,000x reduction from 2004 (Van Dijk *et al.* 2014).

At a smaller scale, the advances and reduction in costs associated with sequencing have revolutionised the study of bacterial genomics. The first bacterial genome was sequenced in 1995. At 1.8Mb in length, the resolution of the *Haemophilus influenzae* was a culmination of efforts – it took eight technicians three months to perform the necessary sequencing reactions, with an estimated labour-inclusive cost in excess of $85,000 (Fleischmann *et al.* 1995). Today – based on the cost of reagents and consumables at the time of writing - a single technician could sequence *H. influenzae* on a single benchtop Illumina MiSeq for less than £5,000, which would take just 27 hours to run. Additionally, the ligation of unique sequence elements to genomic libraries allows several isolates to be sequenced in parallel, creating an economy of scale that reduces the per genome cost exponentially.

These advancements, along with the increased pervasiveness of NGS, have greatly improved the ability to rapidly identify and track potential disease outbreaks. Using comparative genomic techniques, sequences can be mapped against a common reference or ancestor in order to identify diverging populations and infer epidemiological source. These techniques proved to be of particular importance during an investigation of microbiological terrorism.

In 2001, American authorities recovered several envelopes containing spores of *Bacillus anthracis* – the causative agent of anthrax (Rasko *et al.* 2011). Forensic genomic investigation revealed a high degree of homology between the isolates, suggesting a single common source, and also identified three unique single nucleotide polymorphisms, with phenotypic effects linked to sporulation. These polymorphisms were not present amongst common environmental isolates, but were linked to cultures stored in the FBI repository, allowing the source of the threat to be traced. As a species, *B. anthracis* exhibits a high degree of genetic homology, and due to the spore-forming nature of the organism long periods of genetic dormancy further limit the variation observed (Read *et al.* 2002). However, the successes of comparative genomics aren't limited to the study of genetically invariable organisms.

In 2011, a nosocomial outbreak of carbapenem resistant *Klebsiella pneumoniae* was reported amongst 18 intensive care patients, 11 of which died as a result of the infection (Snitkin *et al.* 2012). Pairwise comparison of WGS data from the outbreak isolates revealed three key transmission events from a single patient, and identified an average of 12 SNPs between each subsequent isolate. The number of polymorphisms were typically higher in isolates that were collected later during the course of the outbreak, suggesting an accumulation of variants - potentially as a result of transmission and pathogenesis.

Also in 2011, two concurrent outbreaks of bloody diarrhoea occurred in Europe, caused by the O104:H4 serotype of *E. coli*. In Germany, roughly 4000 cases were identified, with a second smaller outbreak of 15 cases

occurring in France (Grad *et al.* 2012). The outbreak was traced to a shipment of Fenugreek seeds that was received in Germany during 2009, and later distributed to a German sprout farm and French garden stores (EFSA 2011b). By conventional molecular characterisation (*e.g.* MLST, serotyping) the isolates appeared to be indistinguishable from the ancestral strain. When analysed by mapping genomic reads against the ancestral reference, isolates collected from Germany showed very little variation (2 SNPs). However, isolates collected from the France outbreak showed a much higher level of diversity (19 SNPs). This is hypothesised to be due either to a bottleneck in the German isolates resulting in a purge of diversity, variation in the mutation rates of the individual *E. coli* populations, or an uneven distribution of population diversity in the initial shipment of seeds (EFSA 2011b). The findings from this study highlight that variation can be observed in the levels of diversity present amongst isolates derived from a single infectious source.

3.2.1. Comparative genomics of *Campylobacter spp.*

The majority of *Campylobacter* infections are thought to be sporadic in nature and therefore do not share a single, commonly identifiable source (Cody *et al.* 2012, Olson *et al.* 2008). The sheer volume of cases that occur make routine genotyping infeasible for clinical and public health laboratories, and infections are simply reported nationally at the genus level. Very few publications have used whole-genome sequencing to identify isolates from a common source-point outbreak, perhaps in part due to the difficulties surrounding real-time detection of *Campylobacter* outbreaks and the low discriminatory power of routine clinical tests (*i.e.* basic identification of faecal cultures).

The most comprehensive genetic study of campylobacteriosis analysed WGS data from 384 *Campylobacter spp.* isolates, collected from a clinical microbiology laboratory over a period of four months (Cody *et al.* 2013). The study aimed to describe a quick and comparable method of detecting epidemiologically related cases, that would also be viable for use by non-specialised clinicians. Irrespective of knowledge or training, several factors impact the viability of comprehensive bacterial genomics

in a clinical environment. Both the hands-on and compute-times required for large scale analyses are significant barriers, which typically require high-end (*i.e.* expensive) workstations. Additionally, raw sequence data has a large digital footprint, with next generation sequencing outputting tens of gigabytes of data per run. This makes short and long term storage of data, as well as timely analysis in a confidential and clinically accountable manner particularly difficult (Fricke & Rasko 2013).

The method described within the study – known as whole genome MLST (wgMLST) - addresses many of these concerns, relying on the upload and analysis of *de novo* assemblies, which are orders or magnitude smaller than raw reads, and are easily obtained via automatable pipelines. wgMLST is functionally identical to the MLST methodology described in Chapter 1, whereby loci are assigned an arbitrary allele number based on iterative differences in their constituent nucleotide sequences. Query sequences are then compared against a central repository containing previously identified alleles, providing a whole genome allelic profile of identified loci, and allowing pairwise comparison of other isolates processed in the same manner.

The study considers isolates to be linked if they differ by ≤20 loci. Excluding data from repeat patient sampling, and one instance of likely laboratory contamination, no epidemiological case linkage was detected. This does not mean that case linkage did not occur, as discussed previously, *Campylobacter* incidence is severely under reported (1 in 9.3 cases reported nationally), therefore not all cases would be processed by the clinical laboratory. Additionally, the cut off value may not be an adequate or finite measure for case linkage. Within the study, repeat isolations from a single patient yielded alternative alleles in 3-14 loci, whilst repeated sequencing of some isolates showed an average number of alternative alleles in 1-7 loci. This reflects one of the main limitations of wgMLST, in that the results are purely qualitative, and provide no quantitative measure of variation.

For example, assume locus X is encoded by three nucleotides: ATG. Upon submission to the database, this locus is assigned allele number 1. In another isolate, the same locus has a mutation in the first base: TTG, which is assigned to allele number 2. In a third isolate, all three bases differ from both isolates: GGC, and is assigned to allele number 3. Organisms X and Y and Z are then queried against the database, returning allele numbers of 1, 2 and 3 respectively. It's apparent that the organisms are different, but it's not clear that organism X and Y are more homologous to each other than either are to organism Z.

Additionally, this method only measures variation within CDS that have an equivalent definition in the sequence database. Hypothetical proteins that are inferred from protein homology, but are not present in the current *C. jejuni/C. coli* repository are not analysed. This is demonstrated to affect a dataset from a milk-borne outbreak, during which several novel alleles were identified, but could not be analysed using the wgMLST approach (Revez *et al.* 2014).

Intergenic variation is also undetectable using this methodology. This is of particular importance in *Campylobacter* due to the presence of several phase-variable genes that are mediated by the length of homopolymeric tracts. These tracts contain repeat regions of nucleotides that favour slip-strand mispairing during DNA replication, generating a significant level of diversity both in terms of measurable polymorphisms and subsequent gene expression. Whilst many of these tracts are contained within a phase-variable gene's open reading frame, there are instances of intergenic tracts. Results from the following chapter, as well as research conducted by Revez *et al.* (2013), demonstrate that variation occurs within intergenic poly-G tracts during human infection, which would not be quantifiable via wgMLST.

Revez *et al.* (2014), utilised the same wgMLST methodology described above to compare isolates obtained from a milk-borne outbreak of *C. jejuni*. A total of six isolates were compared to the identified source sequence, yielding between 4-12 variable loci. This study also

investigated the differences at the sequence level, and found a limited number of polymorphisms (15 across the dataset). Whilst this approach is favourable, there are still significant limitations due to the use of assemblies in place of raw genomic reads.

Assemblies are consensus sequences derived from the statistically informed merging of raw sequence reads. For example, short read assembler 'SPAdes' breaks raw reads into multiple sized k-mers and produces a de Bruijn graph to link those k-mers into contiguous sequence data (Bankevich *et al.* 2012, Alkan *et al.* 2011). Reads may cover any given position within a genome multiple times, depending on the size and number of genomes sequenced on a platform, as well as the number and length of the reads (Detailed in Chapter 2), however that coverage is collapsed to 1x during assembly. Read mapping confers a significant increase in the number, accuracy and quality of identified variants, and as such confers a finer scale of resolution to epidemiological investigation (Olson *et al.* 2015, Inouye *et al.* 2014).

This study aimed to investigate case linkage between clinical *Campylobacter* isolates across the greater Nottinghamshire area. Whilst links have not been previously established in similar *Campylobacter* datasets, the analysis relied on partial genomic comparisons (*i.e.* wgMLST). By utilising alternative whole-genome methodologies (*i.e.* read mapping), additional fine-scale associations may be apparent that would be undetectable when relying solely on other methods. This is thought to be the first instance in which a read mapping approach has been used to infer the genetic epidemiology of *Campylobacter spp.*, and aims to evaluate the efficacy and comparability of this technique to other methods.

3.3. Methods

Incidence of *Campylobacter spp.* within the greater Nottinghamshire area was defined using a report generated by Queen's Medical Centre, including any isolate with a valid collection date within 2014, and with an assigned laboratory number. Entries without a corresponding frozen isolate were included in the incidence analysis, but excluded from the sampling process. Full details of the sampling and data generation processes can be found in Chapter 2.

MLST profiles providing data on CC and constituent sequence types were obtained using the PubMLST *C. jejuni/C. coli* allele database hosted on the BigsDB platform (Jolley & Maiden 2010). Whole genome assemblies were analysed using Quast (Gurevich *et al.* 2013) to obtain the N50, total assembly length and number of contigs. Assemblies that were deemed to be of insufficient quality for downstream analysis due to an extremely low N50, notably larger or smaller total genome length, or with an unusually high number of contigs were excluded from subsequent analysis.

The core genome phylogeny of assemblies with sufficient quality was reconstructed using the Parsnp tool from the Harvest software suite (Treangen *et al.* 2014). Due to the constraints of the maximal unique matches index, *C. jejuni* and *C. coli* isolates were processed separately, with the datasets comprised of 142 and 11 isolates respectively. The resulting newick formatted trees were visualised using the Interactive Tree of Life (Letunic & Bork 2016), with colour-bar legends prepared manually in post-editing.

The core genome phylogeny was used to infer case linkage. Isolates that shared the same clade with zero branch length were treated as potentially linked, with additional information such as matching CC and sequence type used to qualify further investigation. Due to the overall phylogenetic distance within the *C. jejuni* dataset, additional phylogenies were constructed for CC's displaying potential case linkages.

Isolates within the intra-complex phylogenies that appeared to be associated were then compared. *De novo* assemblies were analysed using the BigsDB platform to obtain whole genome MLST profiles (Jolley & Maiden 2010). A cut off value of ≤20 variable loci in a pairwise comparison was used to indicate linkage. Loci that were unique to an isolate's assembly were discounted from the comparison.

Isolates that were homologous at the core genome levels were also compared using Bowtie2 to map genomic reads of one isolate to a *de novo* assembly of another (Langmead & Salzberg 2012). Polymorphisms were filtered to only include sites with a base and mapping quality of ≥30, and a minimum depth of 8 using SAMtools and vcftools (Li *et al.* 2009, Danecek *et al.* 2011).

3.4. Results

3.4.1. Incidence

Over the 2014 sampling period, Queen's Medical Centre processed 771 stool samples that tested positive for *Campylobacter spp*. 11 isolates did not have an associated date of collection and as such were excluded from analysis. A total of 760 eligible isolates were collected in 2014, with the greatest incidence occurring in June (n=104). There were two additional peaks in incidence in February and September, as shown in Figure 3.1.

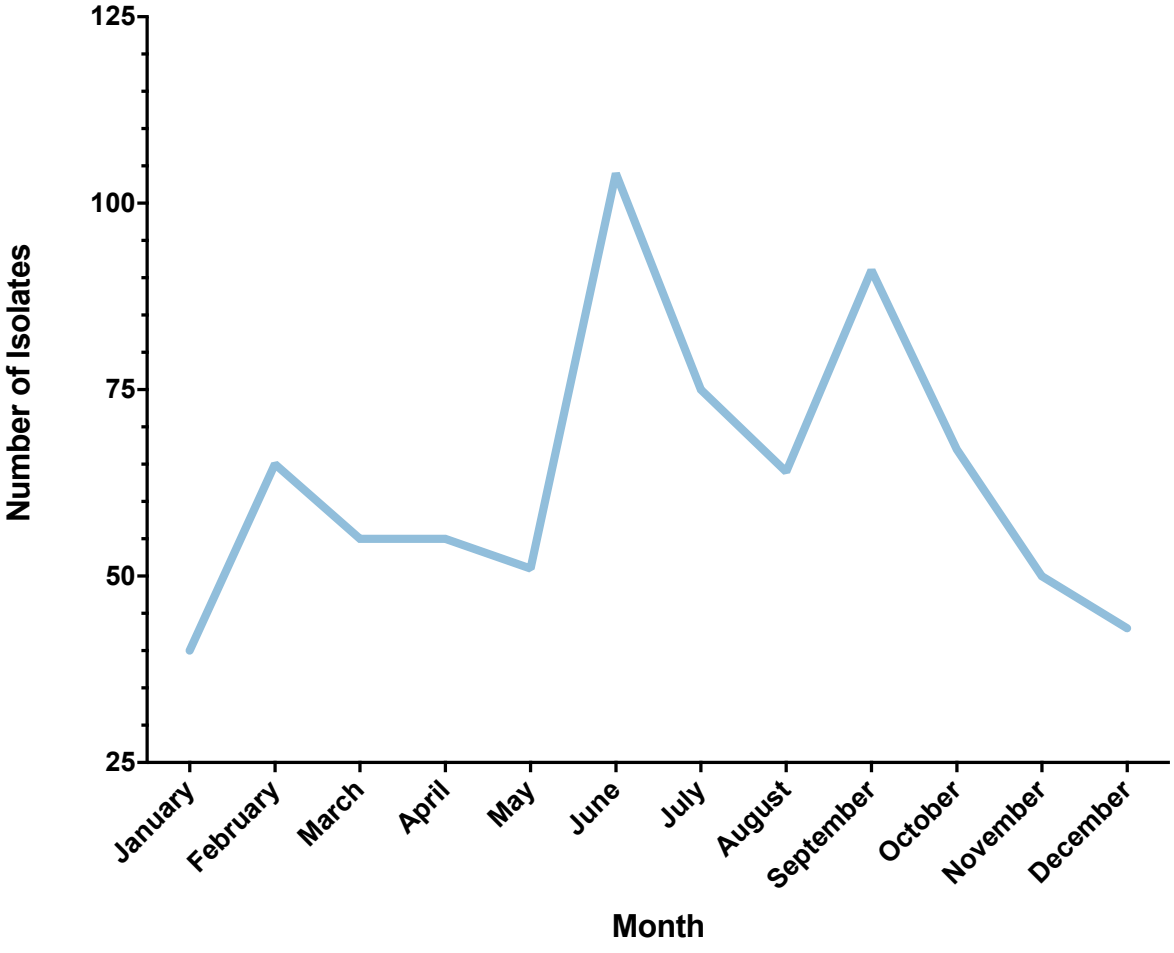

Figure 3.1 - Monthly incidence of *Campylobacter*iosis cases identified across the Queen's Medical Centre's catchment area during 2014

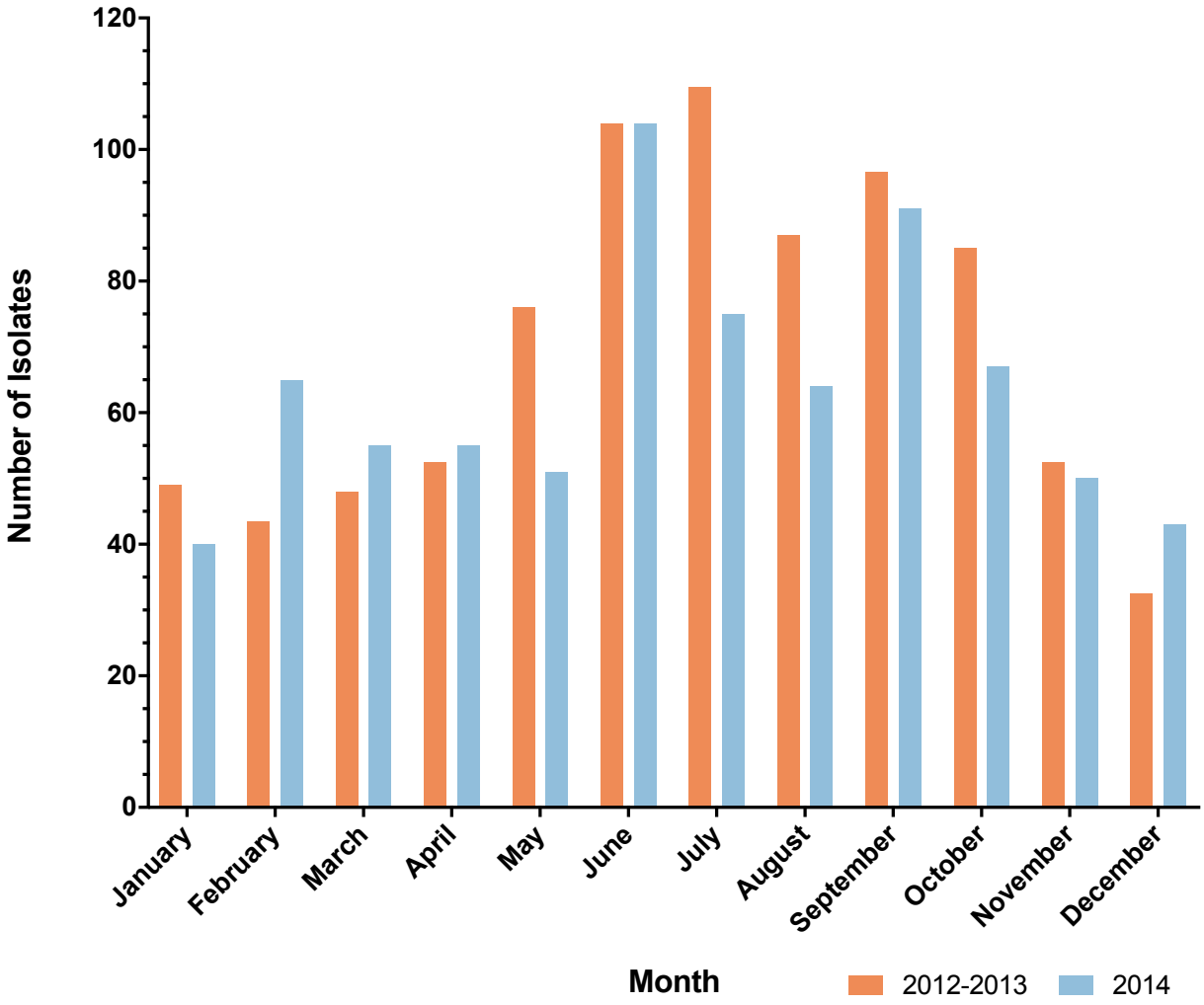

Figure 3.2 - Current vs. historical incidence of *Campylobacter*iosis. within the greater Nottinghamshire area

Data supplied by Queen's Medical Centre, with 'current' incidence constituting data from 2014, and historical incidence formulated form an average of two years' worth of previous data from the same health trust site (data supplied in Appendix 9.2).

Overall, there was a 9.1% reduction in total observed incidence across Nottinghamshire, compared to an average incidence from previous years (~836 ± 14). July and August – typically months with a large incidence – yielded notably fewer *Campylobacter spp*. isolates (75 vs ~110 ±6 and 64 vs ~87 ±0 respectively) as shown in Figure 3.2.

3.4.1.1. February

Cases of *Campylobacter* were higher than expected during February, with a total of 65 cases compared to a local historic average of 43.5. Two 7-day tranches (7th-14th and 21st-28th) accounted for over 73% of all reported infections. These tranches were preceded by two events occurring roughly a week before the increase in cases was observed. Chinese New Year fell on the 31st of January, and Valentine's Day on the 14th of February. Figure 3.3 shows the total daily incidence during February, with both events and 7-day tranches highlighted.

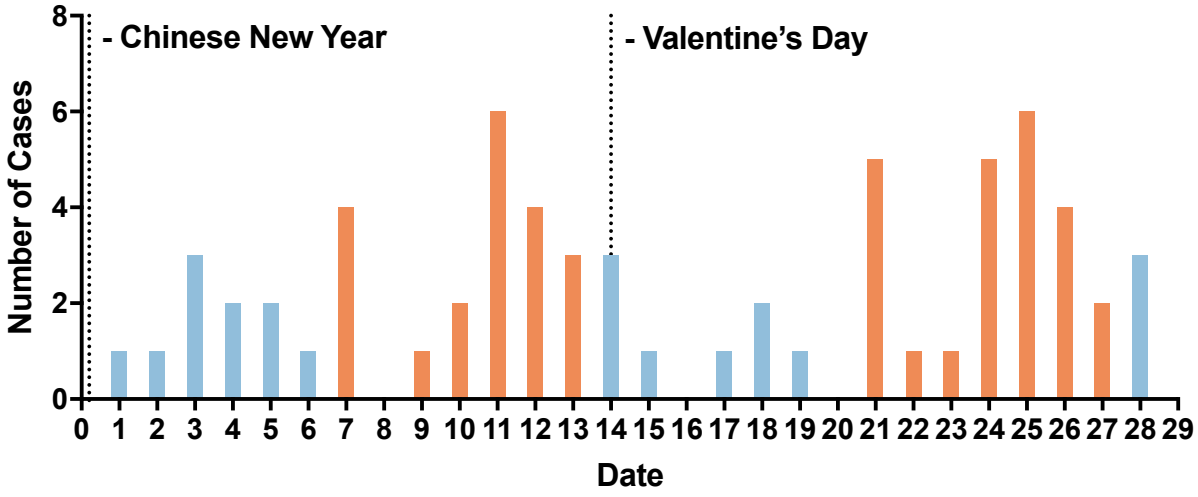

Figure 3.3 - Incidence of campylobacteriosis across Nottinghamshire during February, 2014

Two peaks in incidence occurred following two marked events. Orange bars represent the week long tranches of increased incidence.

Incidence data from the United Kingdom was analysed to measure isolation rates during February from previous years. Figure 3.4 shows monthly *Campylobacter spp*. incidence ranging from 2008-2011, formatted in monthly and daily averages. The latter accounts for the deviation of month length in days, with consideration to leap years. Both sets are normalised to a percentage of the total dataset to allow comparison. When formatted as a daily average, an additional peak in February can be observed, reflecting a national distribution of this previously undescribed peak.

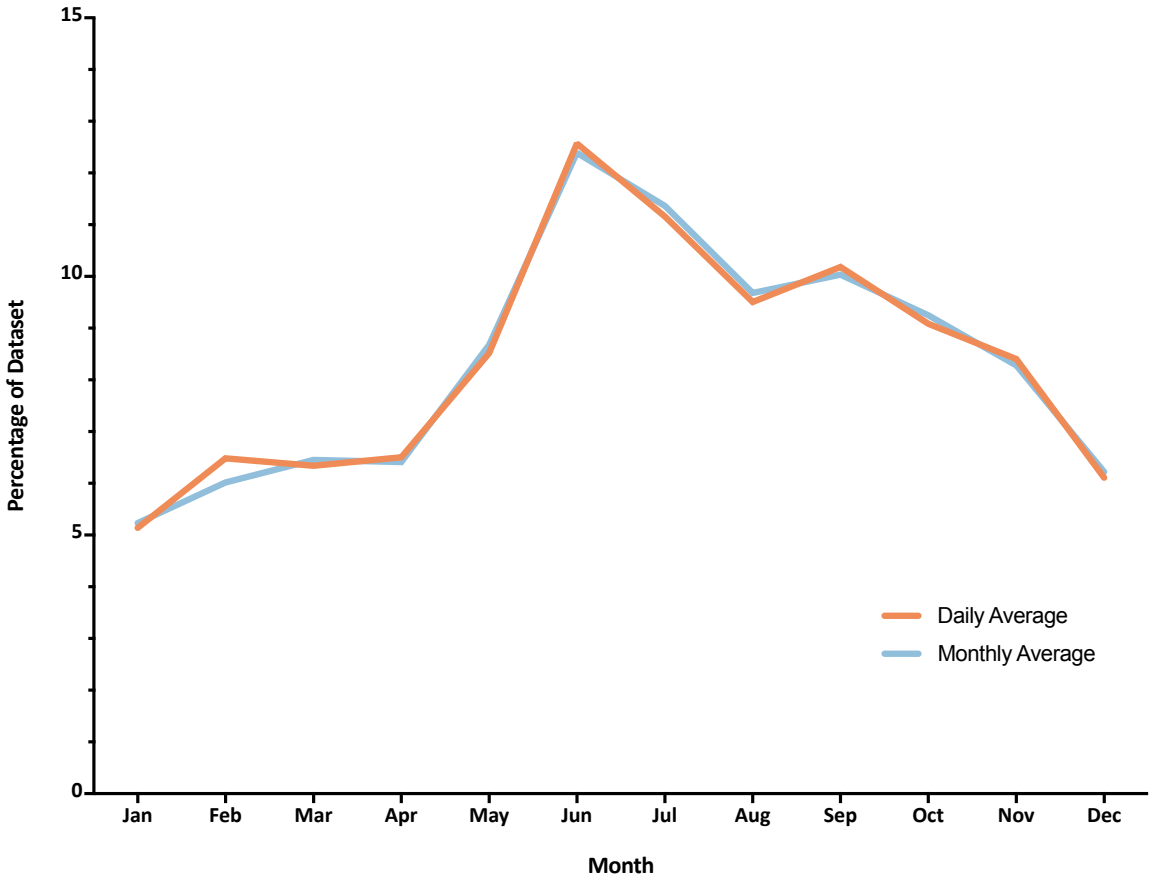

Figure 3.4 - Percentage of average *Campylobacter spp.* incidence across the United Kingdom, from 2008-2011*

* - Data obtained from PHE (DEFRA & PHE 2015).
Normalised to account for the length of the month in days.

Additional metadata encompassing patient age and sex was available for 766 isolates. Across 2014, 55.7% of cases were isolated from males. Figure 3.5 displays the monthly distribution of cases amongst age groups. This shows that for the majority of groups, the peak in age-specific incidence coincides with the seasonal peak observed in the summer. Certain age groups exhibit raised incidence outside of the seasonal peak (*e.g.* 5-10 year olds had the highest incidence during April).

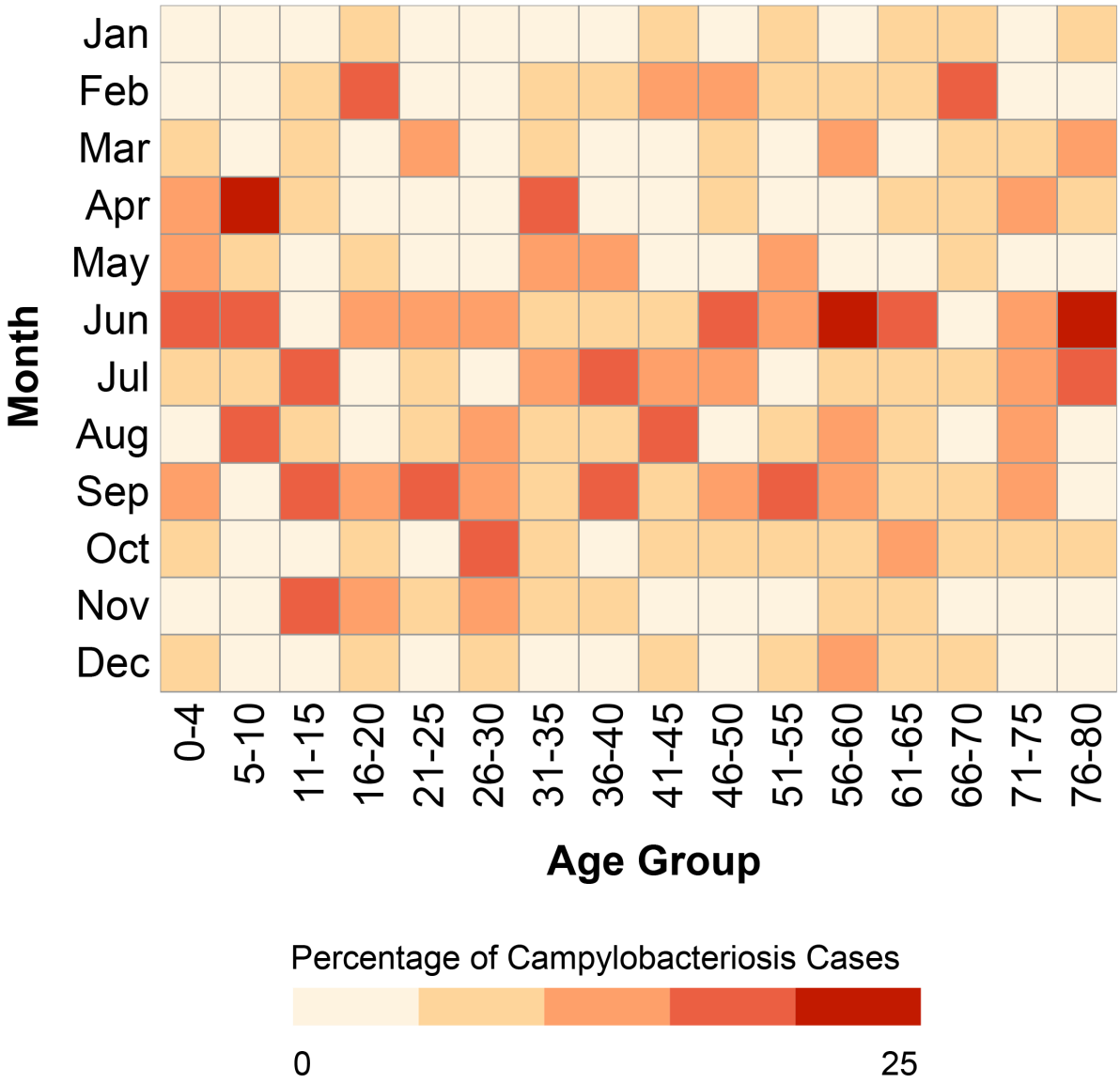

Figure 3.5 – Monthly distribution of campylobacteriosis cases amongst age groups in the greater Nottinghamshire area

In a similar fashion, Figure 3.6 shows the distribution of constituent age groups of campylobacteriosis cases by month (*i.e.* percentage of cases within a given month attributed to a single age group). The majority of months follow a typical age distribution, with the bulk of constituent age groups lying between the 16-65 marks. February is of particular interest as no cases occurred in persons under 15 years, or over 70 years of age.

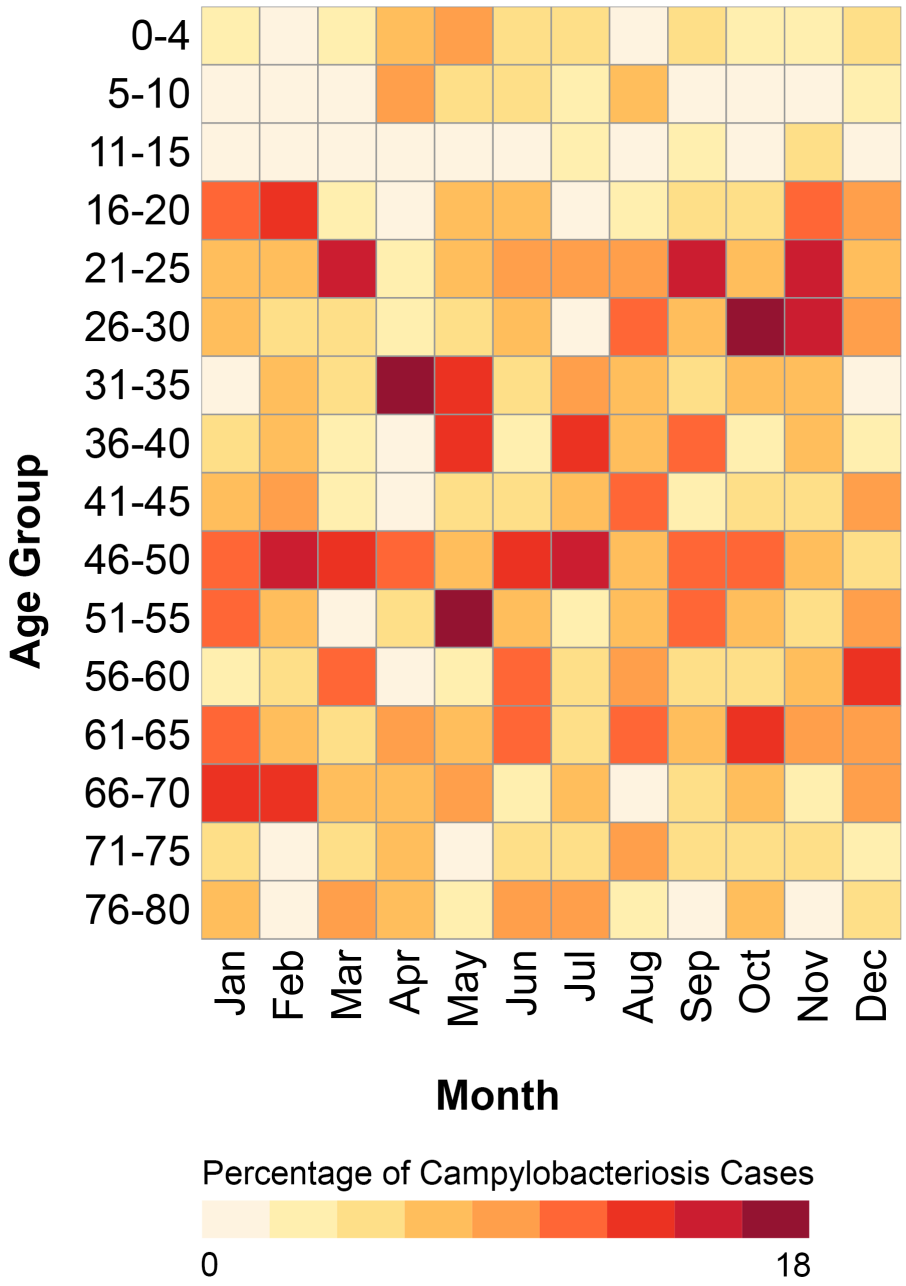

Figure 3.6 - Monthly incidence of campylobacteriosis across the greater Nottinghamshire area with abundance of constituent patient age groups

A total of 664 *Campylobacter* samples were isolated from patients living within the greater Nottinghamshire area. Figure 3.7 provides a heat map of *Campylobacter* infections assigned to each of the postal boundaries within Nottinghamshire, based on the incidence of campylobacteriosis per 100,000 persons. The highest areas of incidence lie within postal boundaries close to the city centre, and reflect a relatively dense and largely urbanised environment.

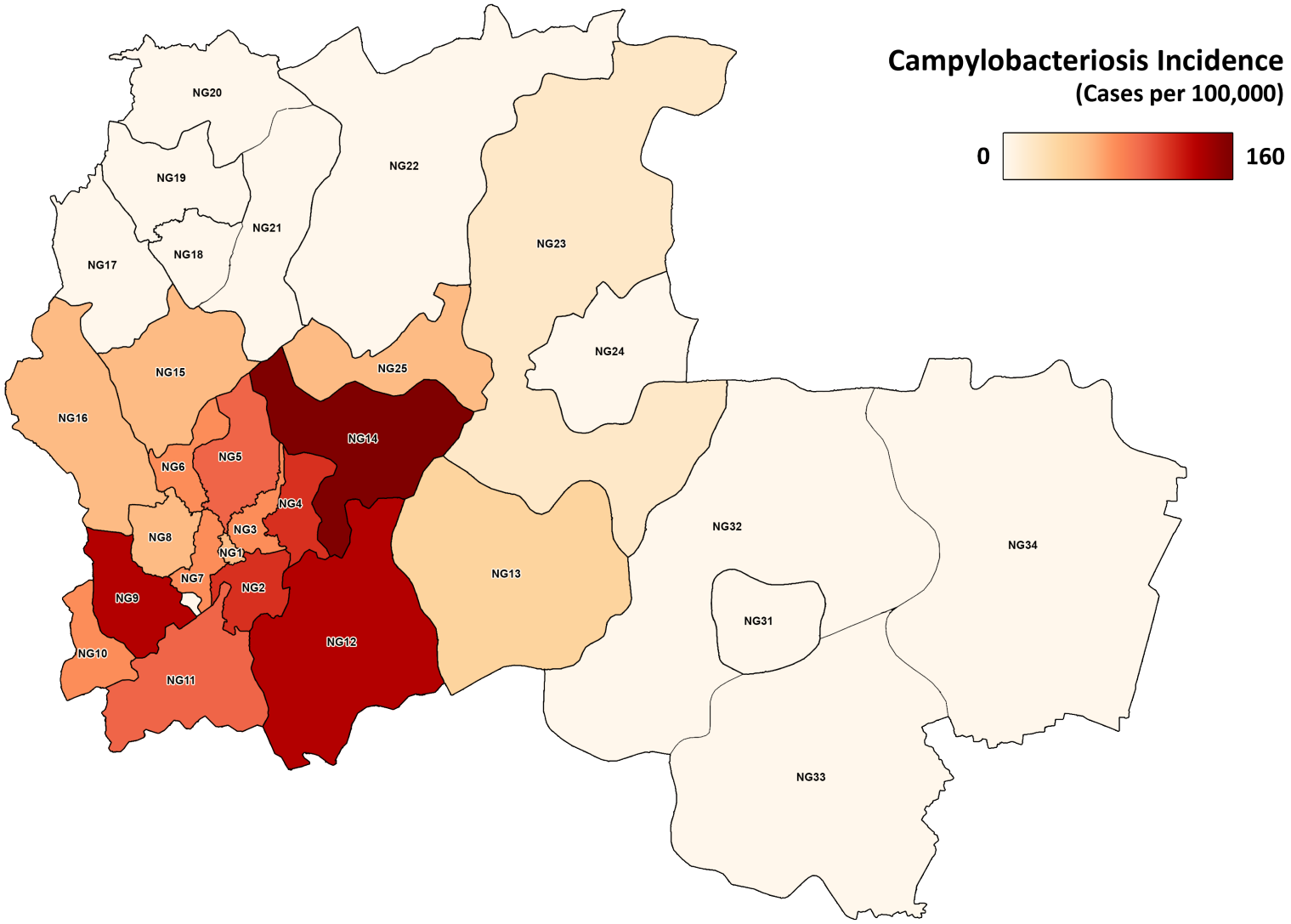

Figure 3.7 - Distribution of *Campylobacter*iosis incidence* across Nottinghamshire postcode areas during 2014

Incidence given in cases per 100,000 people, determined by 2011 population census data (ONS, 2011).

3.4.2. Sequencing and Assembly Quality

A total of 185 clinical *Campylobacter spp*. isolates were selected for whole-genome sequencing. All isolates were collected throughout 2014, and represented 24% of the total number of isolates that were processed by Queen's Medical Centre that year. Prior to further analysis, the generated sequence data was assessed for quality, with contaminated or otherwise inadequate files excluded.

The final curated genomic dataset (n=158) had a mean N50 of 120,954bp (95% CI: 96,616bp to 145,291bp), a mean total length of 168,001bp (95% CI: 1,666,466bp to 1,693,696bp) and a mean number of contigs of 169 (95% CI: 129.4 to 208.6). Figure 3.8 shows the distribution of these quality metrics.

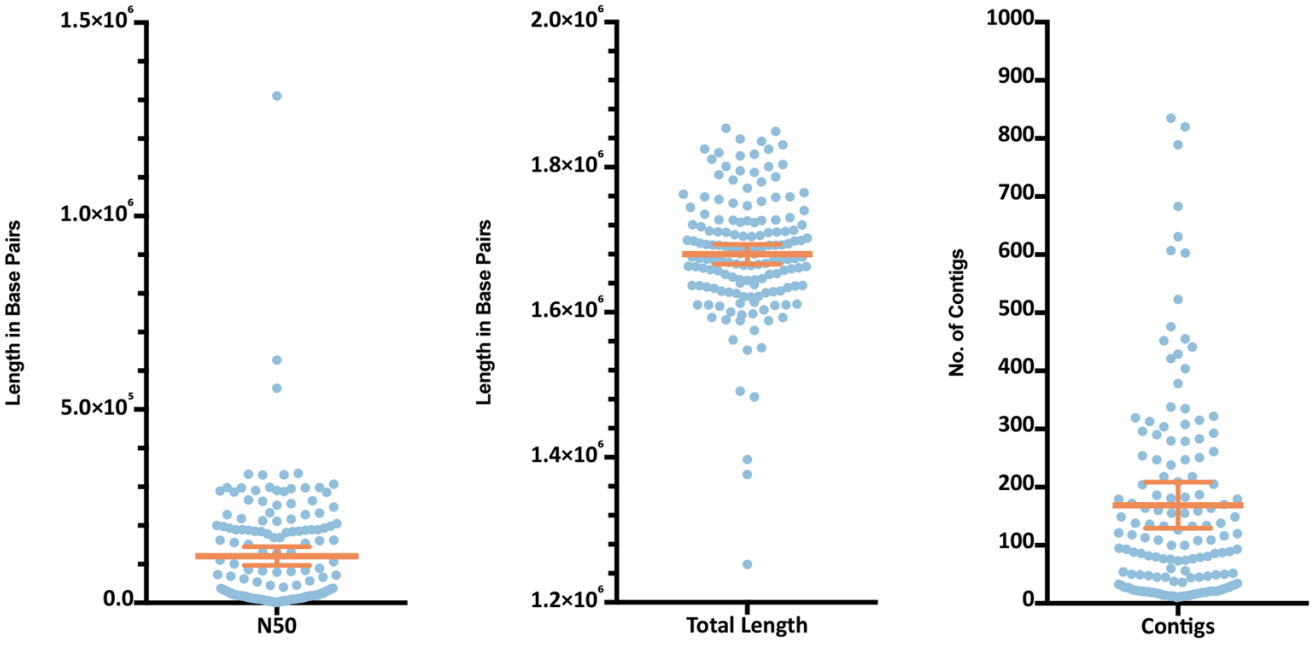

Figure 3.8 – Distribution of quality metrics amongst the final genomic dataset, as reported by Quast

Horizontal bars represent mean values with 95% confidence intervals.

3.4.3. MLST

A total of 158 draft genomes were produced with sufficient quality and coverage depth for use in MLST analysis. Of these 158 isolates, 90% (n=143) belonged to *C. jejuni subsp. jejuni* and a further 7.6% (n=12) were identified as belonging to *C. coli* CC ST-828. Three isolates (1-55, 5-65 and 6-20) yielded novel allele profiles that could not be assigned to an existing CC or ST. Their respective allele profiles are displayed in Table 3.1.

The dataset yielded a total of 29 unique CC's. The dominant complex identified was ST-21, which accounted for 26.5% (n=42) of all isolates. A further 6 complexes constituted more than 5% of the dataset – ST-828, 206, 48, 353, 257 464 and 45. Figure 3.9 shows the distribution of identified CC's.

Table 3.1 - MLST allele profiles for isolates with no identified sequence type

	Isolate		
Locus	1-55	5-65	6-20
aspA	8	2	2
glnA	2	364	28
gltA	2	4	29
glyA	64	38	243
pgm	238	141	286
tkt	3	25	267
uncA	35	35	124

* - locus allele numbers of seven housekeeping genes, as defined by PubMLST scheme for *C. jejuni* and *C. coli*.

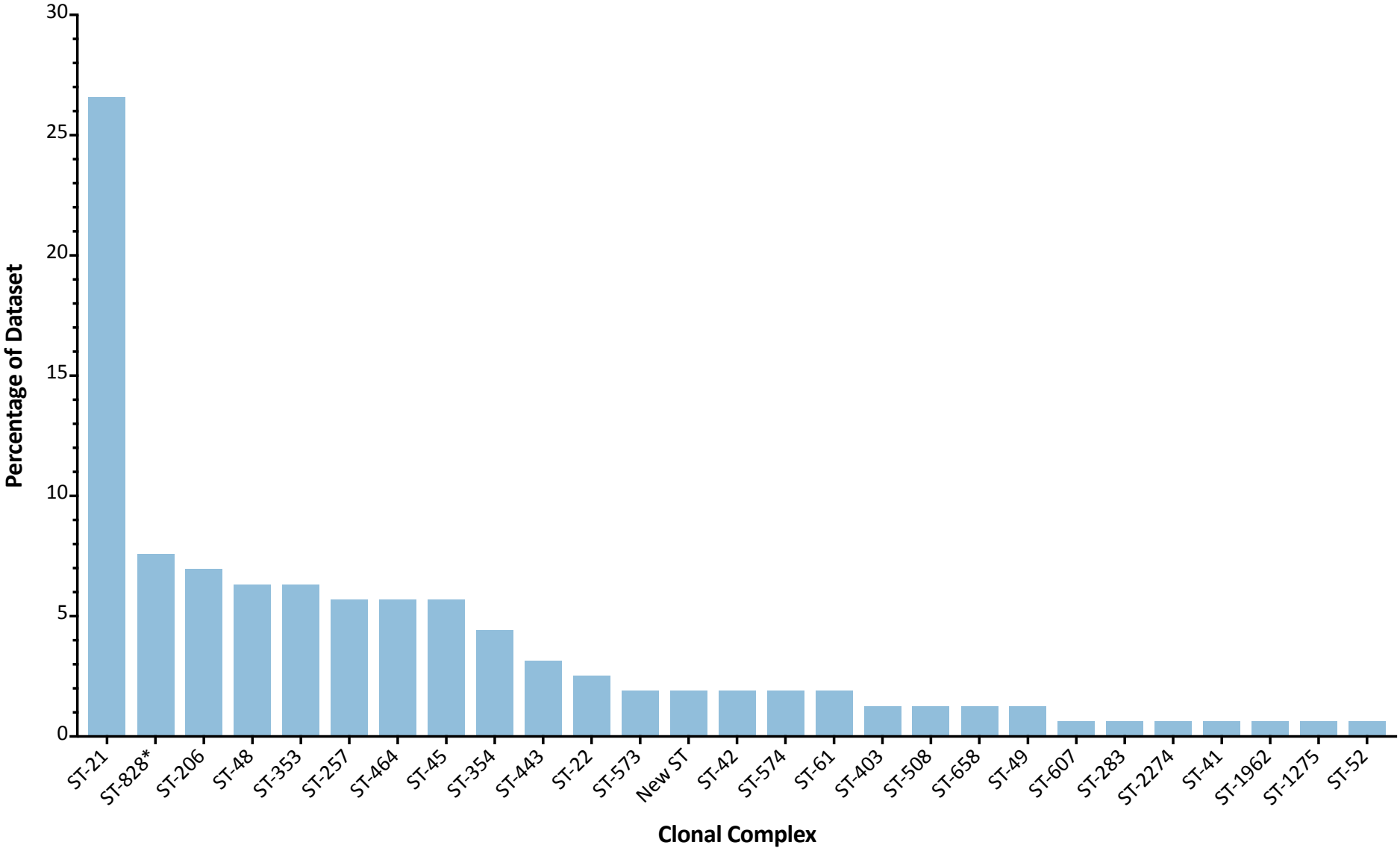

Figure 3.9 - Distribution of clonal complexes amongst 158 clinical *Campylobacter spp.* isolates obtained from Queen's Medical Centre

* - *C. coli* clonal complex.

3.4.4. Phylogeny

Separate core genome phylogenies were constructed for all (n=143) *C. jejuni subsp. jejuni* isolates (Figure 3.10) and *C. coli* isolates (Figure 3.12).

Figure 3.10 shows the core genome phylogeny of all *C. jejuni* isolates. Branches highlighted in red indicate isolates that share a significant genetic homology (*i.e.* branch length: 0) and that were investigated further. This included 18 isolates from the dominant CC: ST-21, which were reanalysed as a subgroup as shown in Figure 3.11.

Figure 3.10 - Core genome phylogeny of 143 clinical *C. jejuni* isolates

Annotated with colour bars representing constituent clonal complex, and red lines indicating isolates that bear significant genetic homology at the core genome level. Rooted to reference sequence 81116. Additional phylogenies are presented below in a clearer fashion.

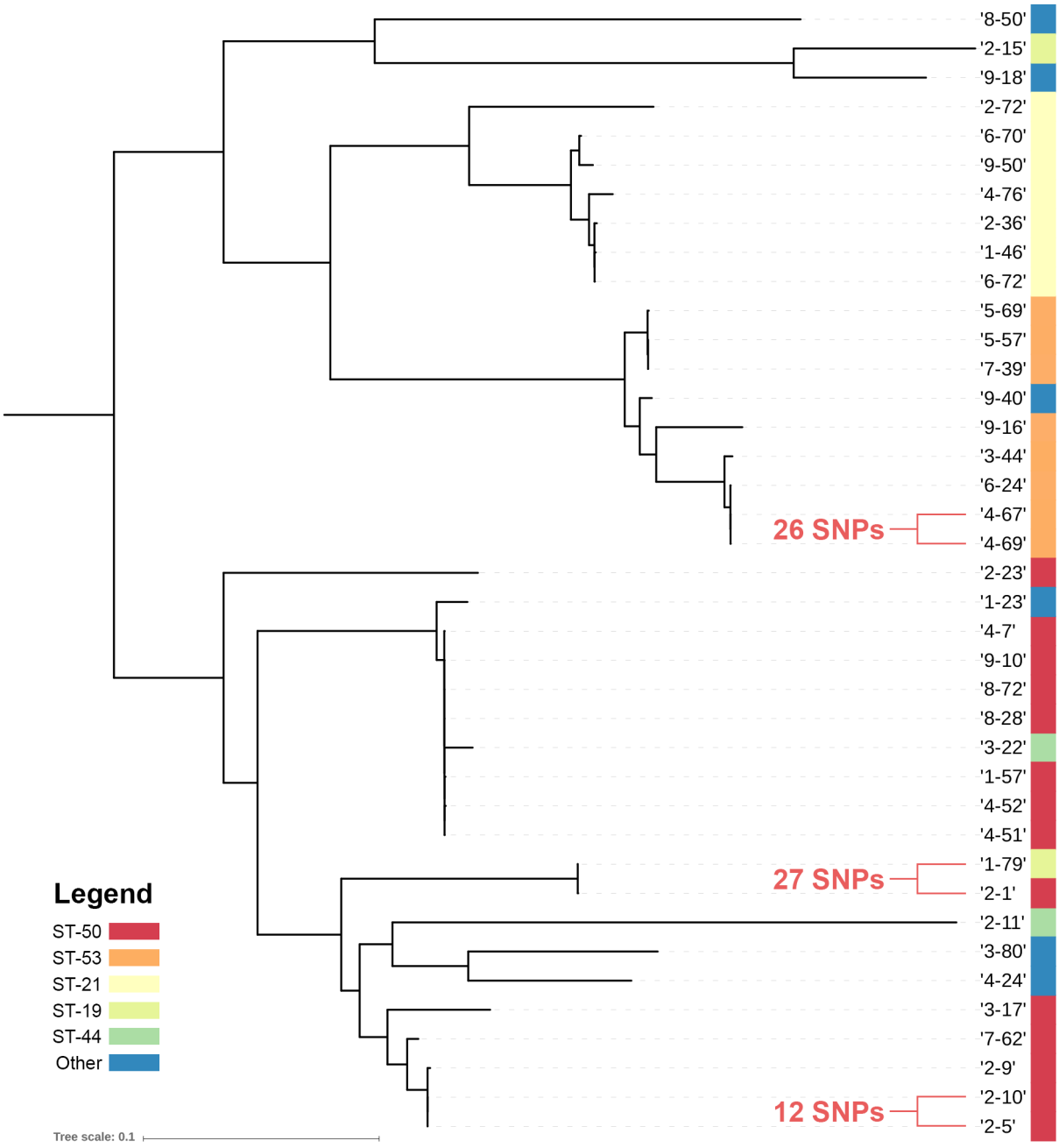

Figure 3.11 - Core genome phylogeny of isolates belonging to clonal complex ST-21 with linked isolate pairs and respective number of Single Nucleotide Polymorphisms (SNPs) highlighted

Tree rooted to mid-point with isolate '9-40' used as a reference. Sequence types with more than one constituent isolate are annotated with colour bars. In instances where a branch length shows as ~0 but is not highlighted, additional analysis revealed diversity outside of the called core region discounting case linkage.

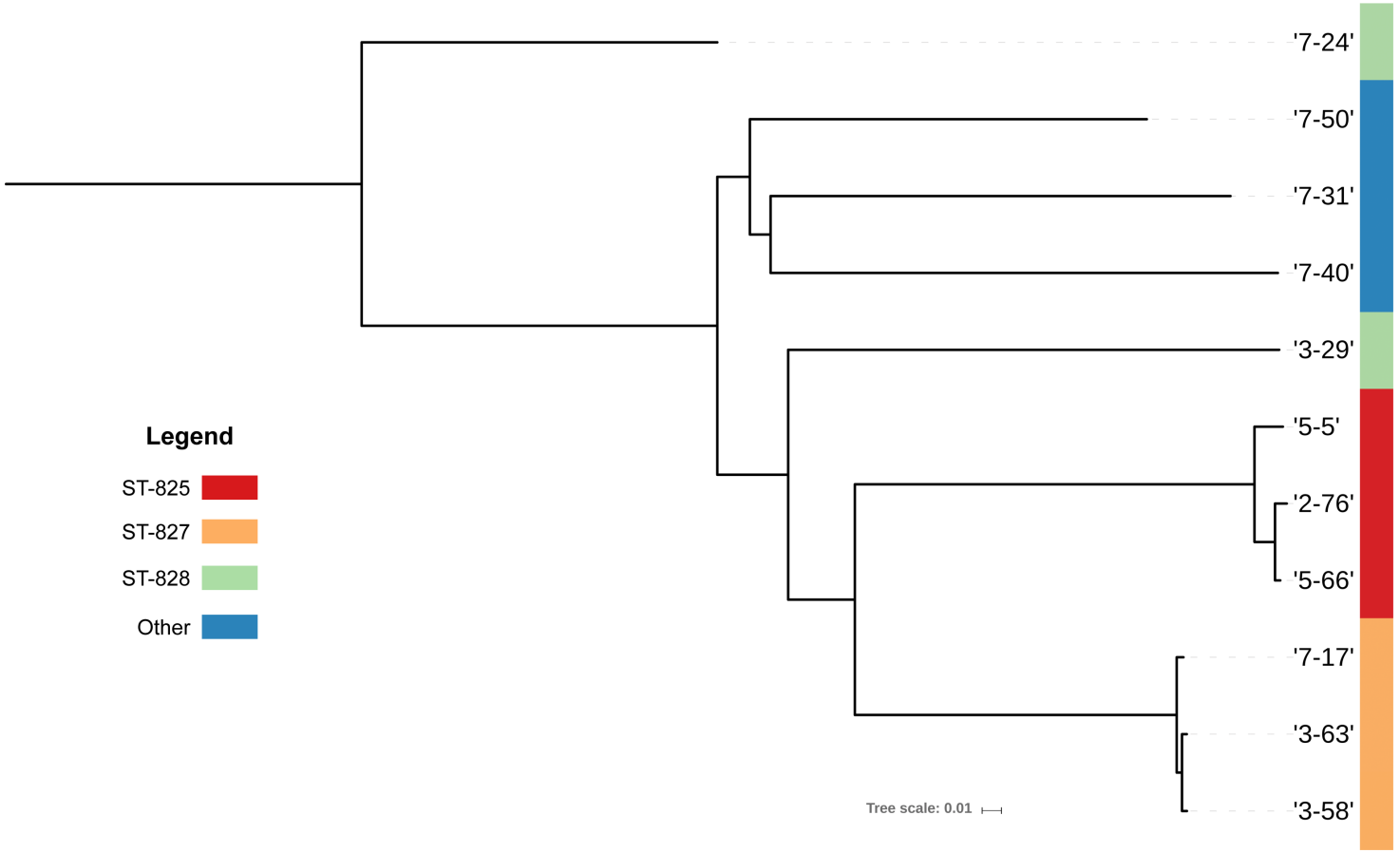

Figure 3.12 - Core genome phylogeny of clinical *C. coli* isolates

Tree rooted to mid-point. Sequence types are annotated with colour bars according to their constituent sequence types

Isolates that appeared to share a common source based on the core genome phylogeny were analysed using wgMLST. Isolates that contained fewer than 20 loci in which a variant allele was identified were then further analysed using alignment to identify single nucleotide polymorphisms (SNPs). A total of 10 isolates were considered to be potentially linked by wgMLST methods (Table 3.2).

Table 3.2 - Homology measured by wgMLST and mapping approaches of potentially linked *Campylobacter spp.* isolates

Isolates*			Clonal Complex	Sequence Type	Variable Loci	Unique Loci	SNPs†
2-5	→	2-10	ST-21	ST-50	1	0	12
4-69	→	4-67	ST-21	ST-53	12	51	26
9-71	→	9-46	ST-354	ST-354	3	6	26
2-1	→	1-79	ST-21	ST-19	2	30	27
4-52	→	4-51	ST-21	ST-53	3	159	34
9-27	→	9-72	ST-464	ST-5136	4	3	47
1-49	→	5-38	ST-353	ST-5	7	116	61
1-46	→	6-72	ST-21	ST-21	17	124	92
4-6	→	9-44	ST-257	ST-257	17	55	142
2-9	→	2-10	ST-21	ST-50	4	235	166

* - Reads from isolates in the right hand column were mapped against assemblies of isolates in the left hand column, based on the isolate that contained the fewest identified loci.
† - Number of Single Nucleotide Polymorphisms with a mapping and base quality of ≥30, depth of ≥8 and present in more than 90% of reads.
Variable loci identified using wgMLST, unique loci are defined as genes present in only one of the two assemblies.

Upon further investigation, four of these isolates contained fewer than 30 SNPs and shared a strong temporal association. The closest relative isolates – 2-5 and 2-10 – yielded 12 SNPs, and were collected a day apart on the 24th and 25th of February respectively. Upon further investigation, it was revealed that these samples were collected from the same patient. Isolates 1-79 and 2-1 shared 27 polymorphisms, and were also collected a day apart during February from separate patients. The isolate pairs – 4-69/4-67 and 9-71/9-46 were also verified as originating from different patients, at 4 and 10 days apart respectively.

3.4.5. Single Patient Isolates

Seven of the 12 SNPs detected in the single-patient isolate pair (2-5 and 2-10) occurred in a single locus – CAMP1252/Cj1335, encoding the membrane protein maf4. The remaining polymorphisms were in a single stretch of intergenic sequence on a short contig with no identified CDS. The point mutation and its effect on the respective amino acid residue is provided in Table 3.3.

Table 3.3 - The position, alteration and effect of detected SNPs to their relative amino acid of a single-patient isolate pair

	Codon			Amino Acid		
Residue	Ref		Alt	Ref		Alt
1644	GG**T**	→	GG**A**	Glycine	→	Glycine
1644	G**G**T	→	G**A**T	Glycine	→	Aspartic Acid
1644	**G**GT	→	**C**GT	Glycine	→	Arginine
1645	CC**C**	→	CC**T**	Proline	→	Proline
1646	CT**A**	→	CT**G**	Leucine	→	Leucine
1647	CT**A**	→	CT**T**	Leucine	→	Leucine
1647	C**T**A	→	C**G**A	Leucine	→	Arginine

Polymorphisms were contained within a single locus – CAMP1252/*Cj1335*.

3.5. Discussion

This study sought to characterise clinical *Campylobacter spp*. isolates obtained from a single health trust site, within the greater Nottinghamshire area. These results reveal that using a read mapping approach to comparative genomics, case linkage can be detected, albeit with relative infrequency.

3.5.1.1. Incidence

A 9.2% reduction in *Campylobacter spp*. isolation rates was observed across Nottinghamshire during 2014, compared to data from previous years. At a national scale, the average annual incidence reported to Public Health England from 2010-2013 was 70,474, with preliminary data from 2014 reporting 70,353 confirmed cases across the UK. This suggests that national incidence during 2014 was unchanged and that the reduction observed is a local occurrence (DEFRA & PHE 2015). This reduction may have several impacting factors, or may simply be a natural fluctuation - there is a lack of sufficient regional data to draw a valid conclusion for the Midlands or Nottinghamshire specifically. National data from 2002-2015 reveals a similar reduction in incidence across England and Wales collectively in 2013 (-9.2% from 2012). Incidence in the years following have not shown any change, and as the only example of such a reduction in numbers it is plausible that the reduction has an underlying cause.

One aspect that could have affected the number of observed campylobacteriosis cases at a local level was the weather. Several studies have shown that weather features such as ambient temperature and rainfall affect *Campylobacter* incidence and poultry contamination levels (Meldrum *et al.* 2005, Tam *et al.* 2006, Bi *et al.* 2008, Nichols *et al.* 2012). Local meteorological data shows the total amount of sunshine hours in 2014 were down from 2013 by ~7.2% (Met Office, 2016). There were also more days with rainfall >1mm (Figure 3.13) which, combined with the decrease in sunshine hours would likely discourage certain risk factors

typically encountered during the summer months. Barbequing has been linked to *Campylobacter* incidence, and has in some instances lead to outbreaks (Studahl & Andersson 2000, Allerberger *et al.* 2003, G. Kapperud *et al.* 1992). In 2014 the UK had an estimated 9.25% reduction in barbecue occasions, which would likely have an impact on the number of cases usually attributed to this source (Kantar Worldpanel, 2015). Additional risk factors such as swimming in natural water sources and contact with environmental isolates during outdoor activities would also likely have been reduced in the poorer weather conditions (Schönberg-Norio *et al.* 2004, Doorduyn *et al.* 2010).

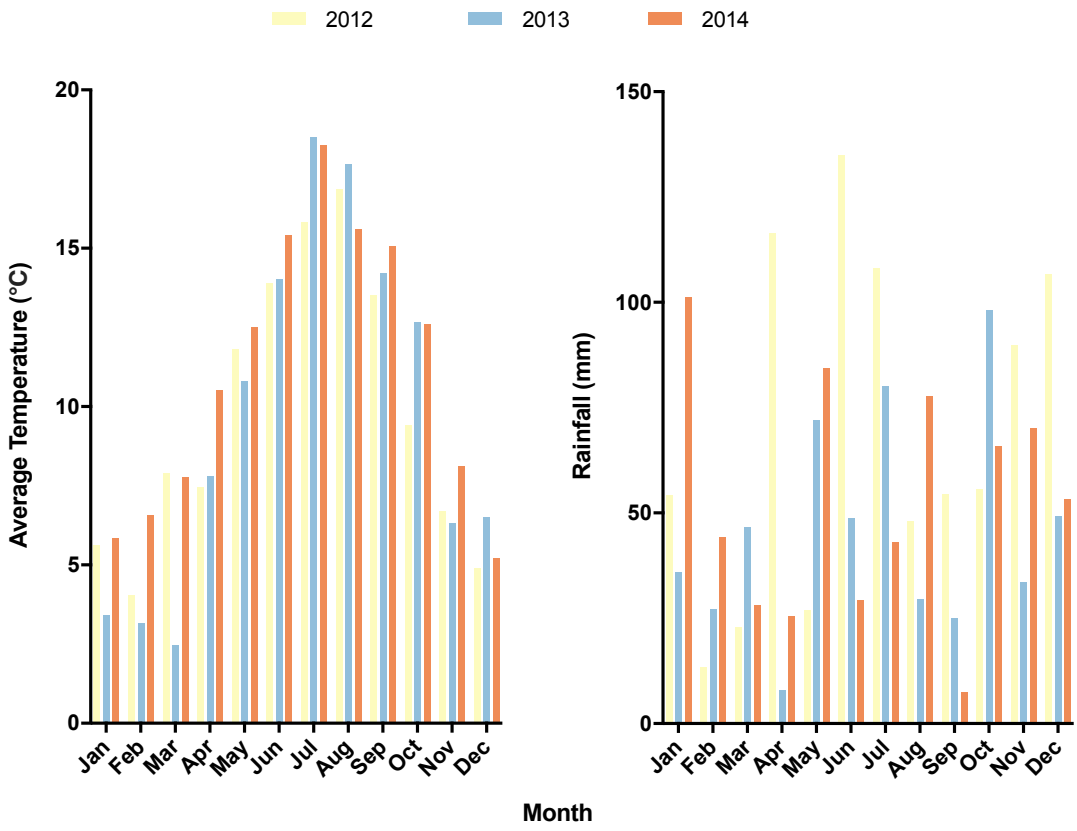

Figure 3.13 - The average monthly temperature and rainfall across Nottinghamshire from 2012-2014

In 2010, the Food Standards Agency (FSA) set a target to reduce total *Campylobacter* prevalence across the food chain, with a particular focus on initially reducing the percentage of samples contaminated with the highest levels (>1000 CFU/g) of *Campylobacter* (FSA & DEFRA 2010). As result of this initiative, several strategies have been implemented, including novel meat processing techniques and large-scale social media

campaigns to aid education and awareness. There has also been a high degree of retailer participation, with several large super markets and vendors participating in the surveillance and reduction strategies. As a result of the study, participating retailers now remove the neck skin from whole carcasses, which was identified as an area that consistently harboured excessive *Campylobacter* loads.

Preliminary data shows that during 2014, a similar level of *Campylobacter* prevalence was observed amongst samples of retail poultry, with 71% of packages found to be colonised. However, there was a reduction of ~8% in samples that were found to contain the highest *Campylobacter* loads of over 1,000 CFU/g, which may have prevented some incidence of campylobacteriosis (FSA, 2010; 2015). The impact of these reductions would likely be more apparent in areas where participating retailers have a strong market presence.

This is particularly relevant to Nottinghamshire, where roughly 85% of the population reside in urban areas, and over 30% in the main Nottingham city area (Office for National Statistics, 2011). Many studies demonstrate that rural areas contribute a higher proportional incidence of *Campylobacter* than urban areas, with fewer cases linked to the consumption of poultry meat and instead associated with animal or contaminated water contact (Nichols *et al.* 2012, Schielke *et al.* 2014, Olson *et al.* 2008).

Given the largely urbanised distribution of cases apparent in Figure 3.7 (*i.e.* higher incidence in densely populated postal areas close to the city centre), it is plausible that the reduction observed in Nottingham may in fact be due to the improvements made in poultry contamination with *Campylobacter*. Further study of the Nottinghamshire area would help to delineate this effect, particularly if there are historic records containing postcode information.

3.5.1.2. Population Structure

The distribution of CC's amongst the sequenced isolates was highly similar to datasets in other parts of the UK. ST-21 is the major constituent CC in the dataset. ST-21 is frequently identified as the main lineage associated with clinical infection, and has been isolated from a wide range of host sources, with many studies also finding it to be the most abundant amongst clinical datasets (Colles *et al.* 2003, Strachan *et al.* 2009, Cody *et al.* 2012). Clonal complex ST-828 – a *C. coli* grouping - is the second largest complex in the dataset. *C. coli* is often attributed to roughly 5-10% of all campylobacteriosis incidence, which is reflected herein by a 7.6% isolation rate (Cody *et al.* 2012, Nichols *et al.* 2012, Olson *et al.* 2008). A further 6 complexes constitute more than 5% of the total dataset and as such are referred to as 'major' complexes, consisting of: ST-206, 48, 353, 257 464 and 45. The distribution of identified major CC's is identical to that of a larger published dataset encompassing isolates from Oxfordshire, Scotland and Northwest England (Cody *et al.* 2013), indicating that the *Campylobacter* population in Nottingham is representative of and comparable to the UK as a whole.

There were an additional 17 complexes considered to be minor due to a relatively low abundance amongst the dataset. Minor variations in isolation frequency were observed when compared to similar datasets (Cody *et al.* 2012, Cody *et al.* 2013, Dingle *et al.* 2002). Given the wide dissemination of *Campylobacter spp.* across a vast host and environmental range, variance in abundance of these minor complexes is perhaps inevitable, and may be resolved (or indeed exacerbated) with a greater sample size. These minor complexes are also found in low abundance in other clinical studies (Cody *et al.* 2012, Dingle *et al.* 2001), perhaps indicating a relatively infrequent exposure to their infectious source. Many of these complexes have been isolated from agricultural samples (Sheppard *et al.* 2014), which would indicate that these isolates are perhaps more adapted to a particular host niche.

Three *C. jejuni* isolates (1-55, 5-65 and 6-20) were identified as novel sequence types. These isolates yielded full, unique allele profiles that did not belong to any assigned sequence type in the *Campylobacter* PubMLST database at the time of query. These isolates are contained within a sub-clade in the core-genome *C. jejuni* phylogeny, with isolates 5-65 and 6-20 forming an additional subclade indicating a greater level of homology at the core genome level. The novel ST sub-clade branches from two isolates belonging to CC ST-573, which has been strongly associated with poultry sources, providing a potential infectious source of these novel sequence types (Sheppard *et al.* 2014).

3.5.1.3. Case linkage

This study presents the first examples of genetic homology amongst confirmed cases of clinical campylobacteriosis. Four isolates were found to contain fewer than 30 SNPs when analysed via read mapping, with one pair of isolates differing by just 12 SNPs. Of these pairings, 3 belonged to CC ST-21, and a single pair to CC ST-354. Whilst these results are in concordance with the notion that the majority of *Campylobacter* infections are sporadic, it demonstrates that case linkage can be observed at relatively minor levels.

As discussed in the introduction to this chapter, previous studies have relied on incomplete genomic analyses that only assess linkage in CDS in a qualitative manner (Cody *et al.* 2013, Revez *et al.* 2014). The strength of wgMLST as a comparative tool lies in its originally described context (*i.e.* speed, simplicity) and is a strong method of first pass analysis, however these results present the importance of additional analysis to delineate false positive associations.

Using the wgMLST criteria alone, six additional isolate-pairs were considered to be linked. These isolates were later discounted on the basis of excess genetic heterogeneity when measured using a read mapping approach. Additional sample metadata may have been able to eliminate the false negative associations implicated by wgMLST. Temporal

association alone discounts 5 of the 6 spurious isolate pairs. However, the final pair (2-9/2-10) were collected on the same day, and vary at just 4 loci – it's only when these isolates are analysed at the read level that their heterogeneity (126 SNPs) is apparent.

The availability of sample metadata such as sample collection date can vary between and even within datasets. Furthermore, relying on temporal association to eliminate false positives would also obfuscate any single-source infections that occur outside of the timeframe considered (*e.g.* frozen reservoir). Read mapping provides an objective measure of pairwise homology to infer case linkage, which is strengthened rather than qualified by additional sample metadata.

There was a disparity between the number of variable loci identified via wgMLST and number of SNPs identified during alignment. One isolate pair exhibited 7 variable loci, suggesting an initial potential linkage. However, these isolates were later found to differ by 67 SNPs. This is in contrast to another pair with 12 variable loci containing only 26 polymorphisms. This reiterates the limitations of the qualitative nature of wgMLST, and suggests that the ≤20 loci criteria is perhaps a weak measure of true case linkage. Additionally, several loci were found to be unique to an individual assembly (*i.e.* only present in one of the assembly pairs). Differences in coverage and per-base quality at any given locus position will affect the *de novo* assembly process, and may result in variable inclusion of a particular sequence region. These regions were excluded from analysis by utilising the assembly with fewer loci as a reference to which reads were mapped. Comparison in these regions could be achieved via resequencing of the isolate pairs.

The isolate pair displaying the greatest genetic homology (2-5/2-10) were collected from the same patient, with samples taken a single day apart. wgMLST analysis revealed just a single variable locus and read mapping yielding 12 SNPs. As patient metadata was provided after the genomic analysis had taken place, the detection of this linkage acts as a validation of the methodology. No other samples within the sequencing data set

were obtained from the same patient, making this instance unique in its circumstance. It is reasonable to assume that a single patient sample would contain the fewest number of polymorphisms, which is in accordance with our results.

On further investigation, 7 of the 12 SNPs were located within the CAMP1252/Cj1335 locus, which encodes motility accessory factor 4. This locus is one of several phase-variable genes within the *Campylobacter* genome and has been shown to come under variation within same-patient samples in other genomic studies (Gundogdu *et al.* 2007, Cody *et al.* 2013). Three of the seven polymorphisms resulted in a non-synonymous alteration, which would ultimately alter protein function and may reflect an ongoing adaptation to the host environment. As the samples were collected a day apart, the variation observed may be a result of sampling an infectious population at different stages or progressions of the disease.

This level of diversity amongst this isolate pair is similar to the that observed amongst *C. jejuni* 11168 isolates following human and mice passage (Jerome *et al.* 2011, Revez *et al.* 2013). It is also comparable to studies investigating epidemiologically related *Klebsiella pneumoniae* isolates, which also yielded an average of 12 SNPs between patient samples from a commonly identified infectious source (Snitkin *et al.* 2012). *K. pneumoniae* is a genetically diverse species (Holt *et al.* 2015), with a weakly clonal population structure similar to that observed in *Campylobacter* (Wang *et al.* 2013).

Whilst considering the cut off value for defining case linkage in terms of number of SNPs, it was important to consider several factors. Revez *et al.* (2013), measured the level of genetic heterogeneity amongst *C. jejuni* NCTC 11168 following an accidental human infection. A total of 7 SNPs were identified in the passaged isolate when mapping against the completed reference genome.

Similarly, genetic comparisons of 11168 passaged in mouse models yielded 23 instances of variation, suggesting that an increase in diversity may be observed during novel host-colonisation (Jerome *et al.* 2011). NCTC 11168 belongs to the dominant and host-generalist clonal lineage ST-21 and has been isolated from a wide range of sources, both clinical and environmental in nature (Cody *et al.* 2013, Sheppard *et al.* 2014). As such, clinical isolates within this complex may exhibit more or less diversity than the previous studies would indicate, based on the source of the aetiological isolate. Similarly, the diversity present in isolates from other CC's that are associated with a particular host niche is not necessarily directly comparable to the 11168 results. In both 11168 studies, a completed reference genome was available to use for comparison. The necessary reliance on draft genomes within this study may also have an impact on the measurable diversity depending on aspects such as sequence quality and total-genome coverage.

During the routine sampling process, approximately 20% of reported cases were sequenced from a mixture of temporally related and unrelated isolates. It was deemed necessary to test whether an increase in sample coverage would reveal any additional linkage that may be undetectable at a 20% coverage. Following a spate of unusual incidence in February, all samples from a 7-day tranche were sequenced. Due to the fact samples were produced from faecal cultures, some isolates contained an excess of contaminants that were not excluded by selective culture. This, combined with some drop out in the sequencing process resulted in a final sample that represented 75% of the tranche.

Two of the four identified linkages occurred during this period. This was proportionate to a 12% case linkage rate during the week long period, vs 2% for the dataset as a whole. It's not clear whether the increase in observed linkage was as a result of the increased proportional sampling of these isolates, or simply an increase in case linkage during this abnormal period of incidence. To address this, an additional dataset covering a comparable proportion of isolates would need to be sampled,

during a time in which no abnormalities (*i.e.* no increased/decreased incidence) are observed.

Several isolates were later re-sequenced to a greater depth for additional analysis in Chapters 4 and 5. When comparing the reads generated from the deep-sequence dataset to the *de novo* assemblies generated herein, the pipeline employed did not identify any SNPs. Analysis of the resulting assemblies via wgMLST did not identify any variable loci, however, several unique loci were identified in all analysed pairs. The lack of SNPs amongst re-sequenced isolates validates that variation is not falsely introduced as an unintended result of the culture, library generation, sequencing or analysis processes.

The case linkage results are met with the same limitations faced by many genetic epidemiology studies, in that it does not provide a complete overview of the disease. Only one in nine cases of campylobacteriosis will be reported at a national level as a result of a confirmed laboratory analysis (Tam *et al.* 2012), indicating that the total number of *Campylobacter* cases across Nottingham during 2014 would be approximately ~7,200. The severity of campylobacteriosis symptoms is often highly variable, driven by factors such as strain pathogenicity and infectious dose (Butzler 2004), with milder manifestations of the disease contributing to the under reported nature of the illness. Even the most comprehensive clinical study would only be able to sample isolates that produced an illness severe enough for a patient to seek medical aid, which may not reflect the degree to which single-source infections occur in the wider environment. Whilst there is no way to mitigate this, it would be particularly interesting to investigate whether any of the reporting patients' friends and family experienced similar symptoms during the course of their illness, potentially indicating a transient infection that did not lead to a formal diagnosis.

3.5.1.4. February

During the sampling process, an unusually high rate of incidence was observed during February, 2014. Two 7-day tranches of incidence were

found to account for 73% of all samples that were isolated that month. Following infection with *Campylobacter spp.*, there is a mean incubation period of 3.2 days, though this can vary with some reports of periods in excess of 7 days (Silva *et al.* 2011, Janssen *et al.* 2008). In 30% of cases, patients experience a prodromal phase that precedes the onset of diarrhoeal symptoms by approximately 24 hours. It's therefore likely that from the point of sample collection, 3-8 days may have passed since the initial infection occurred. Additionally, it's unlikely that patients will seek medical help immediately after developing diarrhoeal symptoms, which may introduce an additional lag between infection and presentation at a primary health site.

When looking at the dates in which incidence had increased, two antecedent events were identified 7 days prior to the initial increase in cases: Chinese New Year and Valentine's day. These events are widely observed in the United Kingdom, and may see an increase in certain risk-related behaviours. For example, people may be more inclined to eat at restaurants, which during popular celebrations and social observances often offer a reduced set menu. An increase in number of patrons, as well as a reduced variety of dishes eaten may increase the likelihood of multiple single-source infections. Additionally, it may also be assumed that a greater number of celebratory meals are prepared at home, where experience with hygiene and culinary methods may vary.

Additional samples were taken from the peak in incidence following Valentine's day, which provided sequence data for 75% of all isolates occurring within this tranche. Pairwise comparison of isolates with a strong core genome homology revealed two instances of suspected case linkage, representing 12% of the total dataset. It is not possible to definitively conclude whether these links were discovered due either to an increase in the relative proportion of single-source infections, or the enhanced sequencing coverage of this period. However, the fact that the larger dataset excluding these results only identified 2 instances across a period of 12 months (even at a lower 20% coverage) would suggest that at

least a partial increase in case linkage was apparent during the post-Valentine's day tranche.

When considering the demographics of patients presenting with campylobacteriosis, February is unique in that almost all of the constituent age groups were between 16 and 70 years old (Figure 3.6). It is reasonable to assume that these age groups are perhaps more likely to be courting or married, and may as a result be exposed to risk factors associated with Valentine's Day. February represented the 16-20 and 66-70 year age groups' highest monthly incidence (Figure 3.5). These are the only examples of an age groups highest monthly incidence falling outside of the previously defined summer peak in incidence.

February contains the fewest number of days, with 28 during regular years and 29 during leap years. Datasets that report incidence in a monthly format do not adequately reflect this difference, or indeed the lesser differences in other month lengths. Given that Valentine's Day occurs on the 14th of February every year, and Chinese New Year often falls within a similar time frame, any peak in incidence due to these occasions – or other variables unique to February, but not unique to Nottinghamshire - would be observable across multiple years at a national level. When standardising national incidence data from 2008-2011, accounting for the month length in days, a third previously undescribed peak is apparent during February. This would suggest that there is in fact a nationally observed factor that is driving an increase in incidence during this month (Figure 3.4). Additional data combining enhanced genetic surveillance with classical epidemiology (*e.g.* patient questionnaires, point-source investigation) would be of particular use in delineating potential causes of the increase in incidence, and provides an opportunity for future work. It would also be of interest to study the incidence of other food-borne pathogens to distinguish whether the increase is specific to *Campylobacter,* or perhaps observable in many food related illnesses.

Chapter Four: Intrapopulation diversity amongst clinical *C. jejuni* isolates

4.1. Abstract

In conditions such as cystic fibrosis, bacterial evolution has been demonstrated to occur within a single patient throughout the course of their disease (Workentine *et al.* 2013, Lieberman *et al.* 2013). Comparative studies have also shown that these infections are comprised of genetically diverse communities, rather than a clonal population of isolates that share a single common haplotype (Lieberman *et al.* 2013). *Campylobacter spp.* is the leading cause of bacterial enteritis worldwide. It exhibits a vast amount of genomic diversity, and is found across numerous hosts and environmental sources. Existing research has revealed a large degree of genetic heterogeneity amongst clinical *Campylobacter* isolates, and there is limited evidence for within-host diversity during repeat sampling. However, due to the technical limitations of existing studies it is unclear as to whether a typical infectious population of *Campylobacter spp.* exhibits genomic diversity, or indeed how diverse a clinical population may be.

This is the first study to use deep sequencing and novel read mapping approaches to detect and quantify intra-population diversity in clinical *Campylobacter spp.*, with findings revealing genetic heterogeneity in the majority of tested samples. 17 clinical ST-21 *C. jejuni* isolates were sequenced to an average depth of 144X, with read mapping revealing 36 instances of minor allele variation, and several cases of repeat variation at the same amino acid residue.

Four polymorphisms were detected within the homopolymeric tract mediating Cj1321 (encoding a putative transferase), with the same tract-position undergoing a G>C substitution and switching off the expression of the Cj1321 locus. The remaining variants were distributed across 12 loci, with the major functional group encoding chemotaxis related genes. Several isolates exhibited minor allele variation in the chemotaxis gene *ccmL*, with two genomic positions undergoing the same nucleotide substitutions. Variations in the *ccmL* locus, as well as other chemotaxis genes were found to lie within methyl-accepting signalling domains

which may produce a phenotype in which transduction in response to specific ligands is modulated. Variation was also detected in two surface exposed outer membrane proteins, which may reflect epitope polymorphism in response to the host-immune response or adhesion.

These results clearly show that clinical populations of *Campylobacter jejuni* ST-21 are genetically heterogeneous entities. However, it is unclear as to whether the observed diversity is a result of human passage, is unique to ST-21, or whether that diversity is apparent prior to clinical infection.

4.2. Introduction

4.2.1. Bacterial evolution

Bacterial evolution is ultimately mediated via alterations to genomic DNA. These changes include gene modification, gene gain and gene loss. Genomic study of bacterial evolution typically involves comparing the whole genome sequence of an ancestral isolate to its progeny (or other suitable reference to a query sequence). The smallest measure of variation is at the per-base nucleotide sequence level – computational analysis can reveal singular changes to genomic base-pair positions known as fixed variants, or single nucleotide polymorphisms (SNPs) (Pallen & Wren 2007). The process of detecting these variant sites is also referred to as SNP calling, SNP analysis and variant calling (Altmann *et al.* 2012). The first instance of SNP analysis was used to sub-type populations of the genetically monomorphic bacteria *Mycobacterium tuberculosis* (M Cristina *et al.* 2005). Historically, measure of this sort of mutation was conducted by analysing variation in protein isoforms, and took considerable time and effort to investigate (Ketley & Konkel 2005). Using next-generation sequence data, it is now possible to infer phylogeny with exponentially greater haste and accuracy.

Variant calling techniques have been used in a variety of applications, such as the forensic investigation into acts of terrorism involving *B. anthracis* described in Chapter 3 (Rasko *et al.* 2011). These techniques are also useful to track the genetic basis for a phenotypic change, and have been used to track *in vitro* evolution on much smaller time scales. Herring *et al.* (2006), noted that the growth of particular *E. coli* strains on glycerol-containing medium accelerated after several rounds of subculture. The researchers subjected 5 strains of *E. coli* to a 44-day 'adaptation period' of repeated growth on glycerol-containing medium and used microarray sequencing to compare the ancestral and progeny populations. 18 PCR-verified SNPs were detected across the five different strains, in addition to a higher observed rate of growth (Herring *et al.* 2006). SNP calling has also been used to observe genetic variation

amongst outbreak isolates of bacterial pathogens. As discussed in Chapter 3, comparative genomic techniques have been used to successfully track outbreaks of organisms such as *K. pneumoniae* and *E. coli* by analysing the distribution of shared polymorphisms amongst isolates (Snitkin *et al.* 2012, Grad *et al.* 2012).

Some species of bacteria form genetically uniform lineages, with very little variation between isolates. *B. anthracis* is a notable example, with distinct isolates typically sharing >99% nucleotide sequence identity (Read *et al.* 2002). Other species however can contain a far greater genomic variability between strains. Comparison of *E. coli* K-12 and enterohemorrhagic strain O157:H7 revealed 1.4-Mb of sequence that was unique to O157, which represents approximately 25% of the total genome length (Hayashi 2001). The degree of genomic uniformity amongst any given species is a result of the combination of mechanisms by which genetic variation occurs (*i.e.* lateral genetic exchange, homologous recombination and spontaneous mutation). The cumulative effect of these mechanisms can be defined collectively as a total mutation rate. The total mutation rate of bacterial pathogens varies greatly, with WGS revealing a range of 0.1 mutations per Mb per year in the genetically monomorphic species *M. tuberculosis* to as much as 19 mutations per Mb per year in the highly diverse *H. pylori* (Kennemann *et al.* 2011, Ford *et al.* 2011). The mutation rate for *Campylobacter* however has not been robustly defined.

4.2.2. Within-host variation

Many studies have demonstrated that infectious populations of common pathogens are genetically diverse entities, and undergo genomic variation during the course of a typical infection (Young *et al.* 2012, Kennemann *et al.* 2011, Mwangi *et al.* 2007, Nguyen & Singh 2006, Ford *et al.* 2011). Bacterial pathogens with a high rate of genomic variation, and that produce a chronic pathology can logically be observed to undergo adaptation within a single host.

Cystic fibrosis is a genetic disorder that causes impaired immune function and reduced mucus clearance, which ultimately leads to bacterial colonisation of the lower respiratory tract. Whilst the lungs have an abundance of organic compounds suitable for microbial growth, bacteria must continually adapt to changes in specific nutrient deficiencies, as well as the routine, prolonged exposure to antimicrobial therapy (Nguyen & Singh 2006). *Pseudomonas aeruginosa* is found in up to 85% of cystic fibrosis sufferers by the time they reach adolescence, and is a significant cause of mortality (Davies 2002). Phenotypic study of *P. aeruginosa* populations isolated from the sputum of cystic fibrosis patients reveals a large degree of heterogeneity, with a range of colony morphologies and antibiotic resistance profiles (Oliver *et al.* 2000). Genomic comparison of *P. aeruginosa* collected from a single patient at different time points also indicates an on-going pattern of genome-wide adaptation, with 68 SNPs detected in a single clone across a 96 month period (Smith *et al.* 2006).

Within-host diversity has also been demonstrated in gastrointestinal disease. *H. pylori* infects as much as half of the global population (Kusters *et al.* 2006). Carriage of *H. pylori* is chronic in nature, and whilst carriage is not in itself a disease, it has been implicated in a number of pathologies ranging from peptic ulcers to gastric cancer (Kusters *et al.* 2006). Kenneman *et al.* (2011), compared *H. pylori* isolates collected from the same patients at varying time points, with intervals ranging from 3 to 16 years. The genomes revealed 27-232 SNPs between isolate pairs, with additional diversity present in areas of apparent recombination (Kennemann *et al.* 2011).

In both examples, collection and subsequent analysis of a single infectious population has occurred in the magnitude of months apart. It is clear that variation at this time scale does indeed occur. The methodologies employed by these studies (*i.e.* comparison of *de novo* assemblies) are only sensitive to mutations that have become (or are close to becoming) fixed within a population, and as such do not provide

an insight into developing mutations, as there is no data concerning the frequency of these alleles.

4.2.3. Limitations of assembly based genomics

Under strong selective pressure, a mutation will move towards fixation amongst a population. This is seen in positive selection, whereby an advantageous mutation leads to an increase in fitness and an eventual dominance of clones that share that mutation ensues. The reverse is also true in negative or purifying selection, wherein a disadvantageous mutation leads to a decrease in fitness, causing a deleterious effect (Pallen & Wren 2007). Measuring bacterial mutations that have swept to fixation is logically (and computationally) simple, and requires just a single read of a reference and query sequence. This can therefore be achieved at the whole genome level using *de novo* assemblies, which represent a consensus contiguous sequence generated from multiple sequencing reads.

Relying on *de novo* assemblies for comparative genomics has its advantages. The files are smaller in size compared to raw sequencing reads, which in turn requires less processing power and memory to analyse, and far less hard drive space to maintain. However, it also limits the sensitivity of computational analyses. Due to the nature of the assembly process, any analysis using *de novo* assemblies alone assumes that the sequenced population shares a single common haplotype. Genetic mutations that occur in the majority of a sampled population will by extension occur in the majority of sequencing reads. These mutations will therefore be represented in the final assembly as either a break in sequence (*i.e.* a new contig) where the number of variations is sufficient (*e.g.* large insertion/deletion or sequencing error events), or a consensus base call that supports the majority of reads at a variable site.

A break in contiguous sequence would obviously prevent comparison at any omitted bases, which may only be avoided by utilising closed, complete genome sequences, which are labour and cost intensive to

produce. In instances where a consensus base call is produced, there is no distinction between a mutation that has become fixed in a population, and a single nucleotide polymorphism that is present in a sufficient percentage of the sequencing reads to produce a consensus base call.

4.2.4. Measuring intra-population diversity

Lieberman *et al.* (2013), reason that there are two conflicting models of within-host bacterial diversity. The 'dominant lineage' model, which suggests that over time, beneficial mutations that confer a greater fitness will drive clones sharing that mutation to fixation. The 'diverse community' model however, suggests that genomic variation is in a constant state of flux, with many lineages rising to intermediate frequencies concurrently. Irrespective of how diversity progresses, prior to fixation in a population, any given mutation must occur at an intermediate or minor frequency. In order to correctly measure intra-populational diversity, the choice of genetic analysis must be sensitive to alleles that are present at these frequencies.

A suitable method for detecting minor frequency alleles is single colony sequencing (Köser *et al.* 2014). In conventional sequencing protocols, DNA is isolated and sequenced from a single isolate treated as a group (*e.g.* whole plate sweep, liquid culture). Single colony sequencing – as its name implies – relies on isolating DNA from multiple individual colonies, and treating each as a separate entity thereafter. By mapping *de novo* assemblies produced from individual colonies to a common reference, polymorphic sites can be identified, and the percentage of colonies in which that polymorphism appears provides the minor allele frequency (maf). This is an indication of the percentage of a population that shares a haplotype, and can be used to distinguish fixed and polymorphic sites. Single colony sequencing is perhaps most useful in instances where there is a measurable phenotype to distinguish variant colonies (*e.g.* antimicrobial susceptibility). In the absence of a distinctive phenotype a large sample size would be necessary to reduce sampling bias, which is both labour and cost intensive (Lieberman *et al.* 2013). The

utility of single colony sequencing is also severely hampered by highly motile organisms that exhibit swarming on solid media, making individual colony isolation difficult.

Another method for assessing intra-populational diversity is referred to as 'deep sequencing'. For most purposes, sequencing an isolate to 20-30x coverage (*i.e.* each position in the genome is sequenced 20-30 times) is sufficient. This level of coverage will typically produce a useable draft genome, and allow basic SNP calling. Deep sequencing uses the same principal, but greatly increases the coverage (100X+) at which the isolate is sequenced. The genomic reads are then mapped against an assembled reference sequence. This approach significantly increases the number of alleles that are compared to the reference sequence. As the technique uses a read mapping approach, each sequenced base is queried an equal number of times to the positional depth. This approach treats the number of reads at a polymorphic site as the total population size, and uses the percentage of reads at that site which support a given allele to infer the frequency at which it occurs. The single colony and deep sequencing methodologies were compared in their ability to discern minor allele variation. Deep sequencing was able to detect alleles that are present in >3% of the population, and is far less cost and labour intensive than the single colony approach (Lieberman *et al.* 2013).

The seminal work by Lieberman *et al.* (2013) provides key insights into within-host bacterial evolution, and addresses many of the questions presented by the previously discussed studies (Kennemann *et al.* 2011, Smith *et al.* 2006). Sputum samples from five cystic fibrosis patients were used to characterise single populations of infective *Burkholderia dolosa* isolates. Using the previously described deep sequencing technique, Lieberman *et al.* discovered a vast amount of diversity in each patient sample. The majority of mutations that were detected did not sweep to fixation. Rather, multiple haplotypes were detected concurrently as described in the diverse community model of within host-diversity.

The concurrent existence of multiple clonal lineages demonstrates that even at a single time point multiple haplotypes can be detected. This raises questions regarding other pathogens and their related pathologies. For example, whether other genera exhibit similar intra-populational diversity, or if the observed diversity is unique to chronic disease types.

4.2.5. The Diversity of *Campylobacter spp.*

As discussed in previous chapters, *Campylobacter* is the leading cause of bacterial enteritis in the developed world. It is also highly genetically diverse, with a weakly clonal population structure that is distributed across a vast range of hosts (Schielke *et al.* 2014, Garénaux 2008, Waldenström *et al.* 2002, Baker *et al.* 1999, Dearlove *et al.* 2015). Many *C. jejuni* isolates are naturally competent, and have been demonstrated to undergo a significant amount of recombination, as well as exhibiting a high mutation rate (Wang & Taylor 1990; Suerbaum *et al.* 2001; Fearnhead *et al.* 2005).

There are a limited number of comprehensive genomic studies into *Campylobacter* diversity, however initial results suggest that *Campylobacter* does exhibit intra-populational variation. Cody *et al.* (2012), identified several variable loci when analysing repeated samples from patients infected with *Campylobacter spp.* Repeat samples were found to vary at 3-14 loci, however information on the variant bases or allele frequency is not available due to the restrictive methodology employed.

There is also a singular study investigating genetic heterogeneity of a human infection with *C. jejuni.* Following an accidental passage of *C. jejuni* isolate NCTC11168, a total of 7 SNPs were detected in the passaged isolate (Revez *et al.* 2013). This represents far fewer variants than previous research investigating passage of 11168 in mice, which yielded 23 instances of variation (Jerome *et al.* 2011). All variants identified in the human passage were located within hypervariable homopolymeric tracts, and no information regarding the allele frequency is provided.

The researchers suggest that human passage – at least of this laboratory strain – does not result in a significant amount of genetic heterogeneity, particularly in comparison to the results from the mice passage.

Whilst this is certainly an interesting case study, it is not possible to extrapolate the findings to real-world manifestations of campylobacteriosis. The infection occurred as a result of an undisclosed accident involving the reference strain NCTC 11168. This frozen stock was provided by the Centre for Disease Control, and is intentionally clonal in nature. Typical clinical infections are unlikely to come from such pristine sources. Using contaminated poultry as an example, any colonising population of *Campylobacter* will have been exposed to multiple selective pressures throughout the slaughter to consumption food chain. Any diversity present at this stage may impact subsequent variation during passage, or in the case of a purified surviving clone, some adaptation would be expected to move from survival through the poultry food chain and entry to the human host. Additionally, the infected person was aware of the accident, and accordingly expected to develop symptoms. This likely altered the typical clinical management of the disease, with intervention and sample collection occurring far sooner than in community based infections.

In this study, we aimed to assess the degree of genetic heterogeneity amongst clinical *Campylobacter spp*. populations. Given that CC ST-21 is the most commonly isolated lineage in human gastrointestinal disease (Cody *et al*. 2012), the study will focus on this grouping as an indication of a successfully pathogenic genotype. This is the largest study utilising a deep sequencing methodology to investigate *Campylobacter* diversity in the context of a single clinical infection, and the first to focus on naturally occurring disease.

4.3. Methods

4.3.1. Sequencing

Isolates belonging to CC ST-21 were selected from the genomic dataset produced in Chapter 3. Supplementary data detailing the patient demographics of analysed samples later revealed that all isolates were obtained from separate patients.

Fresh genomic DNA was extracted from whole plate growth, and sequenced as previously described in Chapter 2. Two alterations were made to the sequencing process in order to increase the coverage depth and quality of the resulting sequence data.

First, a reduction to the total number of samples (N=24) multiplexed on a single flow cell was used. This increases the depth of sequencing coverage at a given position and allows finer resolution of minor allele variation. The number of multiplexed samples was set at the maximum number of isolates that would produce a target coverage of >150 using the MiSeq V3 chemistry, as demonstrated below.

$$MiSeq\ V3\ Chemistry\text{:}\ \frac{250 \times 25,000,000}{1,628,115 \times 96} = 159.95$$

$$\frac{Read\ length \times Data\ output}{Sample\ genome\ length \times No.samples} = Coverage\ Depth$$

Additionally, due to a declining per-base quality towards the tail end (>250bp) of the MiSeq V3 reagent kit's fourth read, the sequencing run was configured to only utilise 250 (of 300 maximum) cycles. Whilst this decreased the overall data output and thus coverage of the run, it served to produce a greater overall number of reads passing the minimum quality filter (Q≥30) and therefore an enhanced degree of accuracy during read mapping. Sequence data was assembled and annotated as previously described in chapter two, using SPAdes and Prokka respectively (Bankevich *et al.* 2012, Seemann 2014).

4.3.2. Read Mapping

Genomic reads were initially processed using the sliding window quality trimmer, Sickle (Joshi & Fass 2011). Sickle was configured to discard reads that were less than the configured sequencing run length (*i.e.* 250bp), and trim reads in which the quality within the sliding 25bp window falls below 20. This sought to maintain only high quality, full length reads for robust variant calling.

The filtered reads were then aligned to their respective *de novo* assembly using Bowtie2 (Langmead & Salzberg 2012). The maximum insert size was raised from the default 500 to 2000 to permit larger paired-end fragment length alignments. Only alignments containing less than 1% of ambiguous characters were included in the output BAM file.

The resulting alignment was processed using the SAMTools suite of tools (Li *et al.* 2009). The final VCF output was configured to only include variants with a mapping and base quality of >30 (*i.e.* >99.99% accuracy), and a minimum depth of 30 reads.

Due to the process of *de novo* assembly, several variants were identified at the breaking points of contigs, any variants that occupied positions close to the start or end of contigs were discarded. Some variants failed to produce the relevant 'INFO' fields necessary for further analysis, which were also discarded.

The curated VCF was visualised against the read-distribution in Artemis (Rutherford *et al.* 2000). This sought to assess the position of variants in respect to the read. Due to the relatively high error rates observed at start and end positions in Illumina sequence reads, any variants that occurred in the majority at such positions were excluded from analysis.

Figure 4.1 provides a summary workflow of the deep-sequencing read mapping approach.

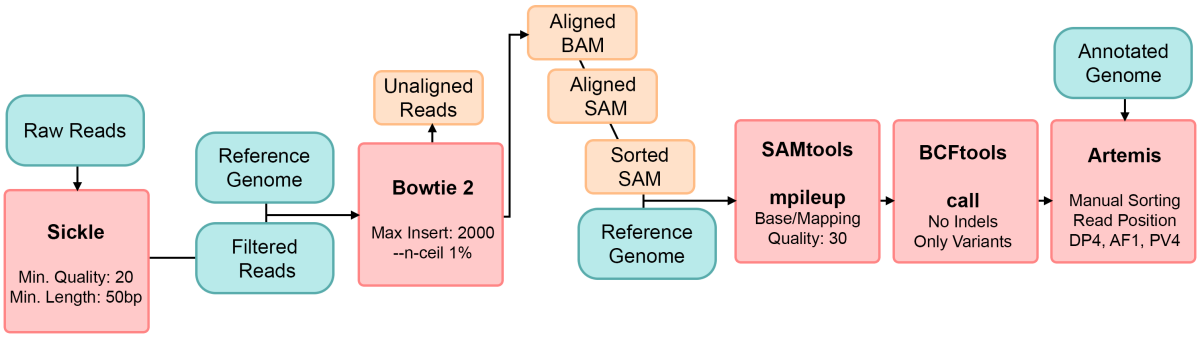

Figure 4.1 - Curation and mapping pipeline for the ST-21 deep sequencing dataset

Software names highlighted within red boxes in bold, with key optional changes to default parameters listed below.

4.3.3. Variant Calls

Annotated assemblies were cross referenced with the genomic positions of called variants to identify potential genes. In instances where variation occurred within CDS, the amino acid residue was used as a reference point rather than raw genomic position, to allow comparison between isolates. Protein sequences were aligned to the Pfam database to identify conserved domains (Finn *et al.* 2016), and - where possible - variants were assigned to a specific functional domain of the protein. Loci are listed primarily by their designation in the PubMLST *Campylobacter jejuni* and *coli* database to allow ease of comparison, and where possible include common aliases thereafter.

The variant call format that provides the final curated list of polymorphic bases includes information on the number of reads that support a given allele, denoted by the tag 'DP4'. There are four fields that comprise DP4: Forward reference (FR) allele, reverse reference allele (RR), forward alternative allele (FA) and reverse alternative allele (RA). This can be used to infer the relative frequency of a minor allele amongst a population of reads using the following equation:

$$Minor\ Allele\ Frequency = \frac{FA + RA}{FR + RR + FA + RA}$$

4.3.4. Statistical Analysis of Assembly Metrics

As query genomic reads were mapped against a *de novo* assembly of the same reads, it was necessary to discern whether various assembly metrics had any impact on the number of minor allele variants detected.

Sequencing coverage depth, assembly N50, and the total number of contigs present in the final assembly were all considered. A linear regression analysis was performed using GraphPad Prism 7, with a significance threshold set at $P>0.05$.

4.4. Results

4.4.1. Isolates, sequencing and assembly.

A total of 17 ST-21 isolates were sequenced to an average depth of 144X. All isolates were found to belong to separate patients, based on age, sex (Table 4.1) and post code (data not shown).

Table 4.1 - Patient demographics of samples belonging to the deep sequencing dataset

Isolate	Collection Month	Patient Sex	Patient Age	Coverage Depth
D1-81	September	M	86	114
D2-11	February	F	79	184
D2-23	February	F	30	103
D2-32	March	F	69	162
D2-36	March	M	57	158
D2-9	March	M	44	146
D3-22	April	F	1	113
D4-24	April	M	23	92
D4-67	April	M	25	182
D4-69	June	M	61	140
D4-7	June	M	47	158
D5-69	June	F	33	169
D5-80	June	M	15	196
D6-57	July	M	63	85
D7-57	September	M	23	160
D7-62	September	M	55	130
D9-10	October	M	21	165

Samples were screened to check whether they were obtained from separate patients. In instances where this was not clear by the provided demographics, geographic data was used for exclusion (data not shown).

The deep sequencing dataset had an average N50 of 384,793bp (95% C.I: 165,003bp to 604,584bp), a mean total length of 1,708,840bp (95% C.I: 1,664,273bp to 1,753,407bp) and a mean number of contigs of 47.12 (95% C.I: 9.345 to 84.89). The distribution of these quality metrics is displayed in Figure 4.2.

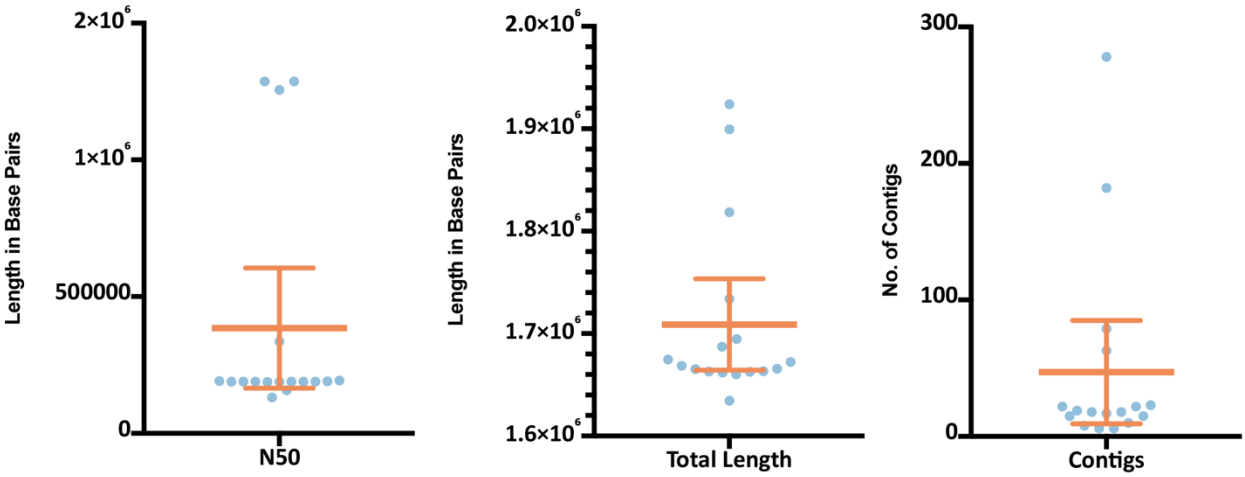

Figure 4.2 - Quality metrics of draft *de novo* assemblies

Mean inclusive of upper and lower confidence intervals highlighted in orange.

Linear regression was used to assess whether the sequencing coverage depth, assembly N50 or number of contigs had an impact on the number of loci in which minor allele variation was detected (Figure 4.3). These analyses revealed no statistical significance (P = 0.6117, 0.2985 and 0.7322 respectively).

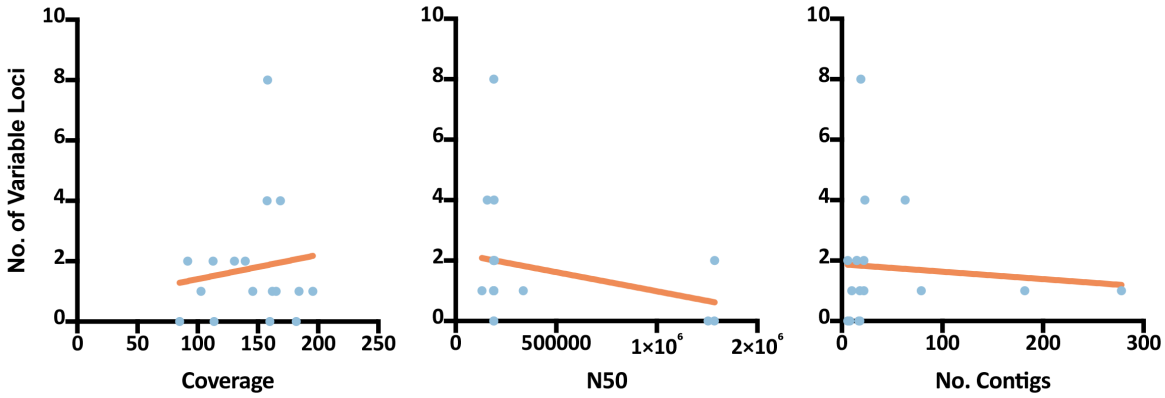

Figure 4.3 - The statistical impact of sequencing and assembly metrics on the number of loci in which minor allele variation was detected

Slopes calculated via linear regression analysis, results were not statistically significant in all instances (P > 0.5).

4.4.2. Minor allele variation

A total of 36 minor allele variants were detected within the dataset with sufficient depth and quality to qualify further analysis. Of the 17 analysed isolates, 35% (n=6) did not produce any minor allele variants that met the strict quality filtering thresholds.

The majority of variants (n=28) were contained within identified loci. Four variants were located within a homopolymeric tract, and a further four variants were present at intergenic positions. These are discussed in greater detail below.

The loci in which minor allele variation was detected were grouped according to the gene product function. Table 4.2 shows the number of isolates (*n.b.* not the number of variants) that contained minor allele variation in these functional groups.

Table 4.2 - Genes in which minor allele variation was detected, grouped according to putative function

Gene Function	Instances of Variation*
Chemotaxis	5
Homopolymeric Tract	4
Intergenic	3
Flagellar	2
Membrane	2

* - *i.e.* number of functionally grouped loci that contained variants in separate isolates. Loci containing multiple variants in a single isolate are counted as one instance. Loci that contain variants in multiple isolates are counted individually.

4.4.3. Chemotaxis

Four isolates yielded variation within chemotaxis related genes. Three of these isolates contained variants within the same gene – CAMP1460 – which is detailed below. The final isolate, D5-69 yielded two minor allele variations in two separate loci (Table 4.3).

Table 4.3 - Minor allele variants detected within loci encoding chemotaxis related proteins

Isolate	Locus	Residue	Codon			Amino Acid			Mapping Quality*	Coverage Depth	MAF†
			Ref		Alt	Ref		Alt			
D5-69	CAMP1031	343	G**G**G	→	G**A**G	Glycine	→	Glutamic Acid	225	428	0.58
D5-69	CAMP1109	275	G**C**T	→	G**T**T	Alanine	→	Valine	46	548	0.2
D3-22	CAMP1460	461	**G**AT	→	**A**AT	Aspartic Acid	→	Asparagine	225	589	0.53
D2-36	CAMP1460	477	**T**TA	→	**C**TA	Leucine	→	Leucine	164	1788	0.28
D2-36	CAMP1460	516	A**T**G	→	A**C**G	Methionine	→	Threonine	157	1840	0.52
D4-69	CAMP1460	523	**A**CT	→	**G**CT	Threonine	→	Alanine	122	763	0.28
D2-36	CAMP1460	523	**A**CT	→	**C**CT	Threonine	→	Proline	225	1838	0.53
D4-69	CAMP1460	550	C**T**A	→	C**C**A	Leucine	→	Proline	225	966	0.30
D2-36	CAMP1460	550	C**T**A	→	C**C**A	Leucine	→	Proline	225	1682	0.52
D2-36	CAMP1460	568	**T**TT	→	**C**TT	Phenylalanine	→	Leucine	225	1476	0.52
D4-69	CAMP1460	615	CA**T**	→	CA**C**	Histidine	→	Histidine	198	691	0.28
D4-69	CAMP1460	621	**A**CT	→	**G**CT	Threonine	→	Alanine	225	612	0.27

* - Provided as a PHRED score

† – Minor allele frequency (*i.e.* indicative of percentage of reads that support the minor allele).

Colours represent sequence type.

Legend: ST-53 ST-44 ST-21

CAMP1031, also referred to as Cj110c, encodes transducer like protein (Tlp) 8, CetZ. Residue number 343 contained a minor frequency variant in the second codon position, causing a non-synonymous mutation from glycine to glutamic acid. This residue is located within the methyl-accepting signalling domain of the protein, as shown in Figure 4.4.

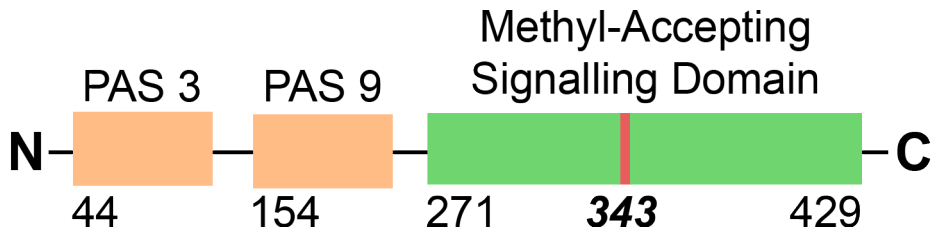

Figure 4.4 - Topology of Tlp8 (CetZ) encoded by the CAMP1031/Cj1110c locus

Residue numbers provided below domains show regions of Pfam alignment. Residue 343 yielded variation at its second base position.
Per-ARNT-Sim (PAS).

The locus CAMP1109 (Cj1190c) encodes Tlp9 (CetA), and again contained a minor variant within the methyl-acepting signalling domain located within this protein. Figure 4.5 shows the topology of the *cetA* locus and the position of the variant amino acid residue.

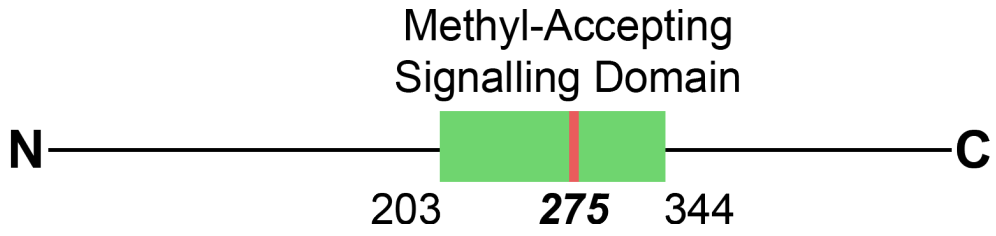

Figure 4.5 - Topology of CetA encoded by the CAMP1109/Cj1190c locus

Residue numbers provided below domains show regions of Pfam alignment. Residue 275 yielded variation at its second base position.

CAMP1460 – also referred to as Cj1564 or *ccmL* – encodes transducer like protein 3 (Tlp3). The majority minor allele variants were contained within the methyl-accepting signalling domain, as shown in Figure 4.6. Residue 523 underwent non-synonymous mutations in two isolates, from the polar amino acid threonine to separate non-polar residues (proline and alanine). Additionally, residue 550 underwent an identical non-synonymous mutation in two isolates, substituting leucine with proline.

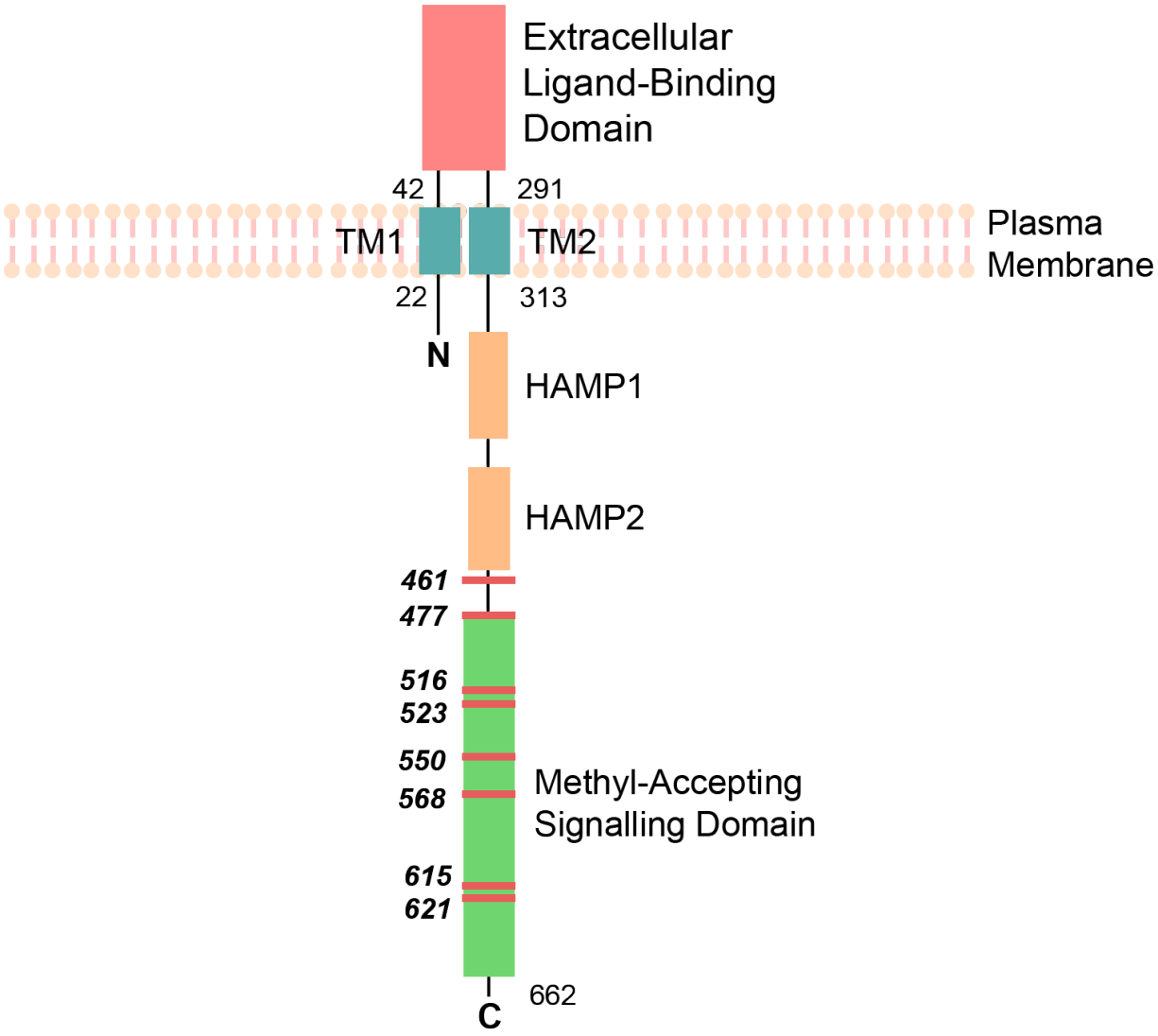

Figure 4.6- Topology of CAMP1460 (CcmL) highlighted with positions of variable residues.

TM1 and 2 = **t**rans**m**embrane helixes.
HAMP1 and 2 = **h**istidine kinases, **a**denylyl cyclases, **M**CPs and **p**hosphatases.
Residue positions provided in bold italicised font adjacent to red lines correspond to minor allele variations detected and listed in Table 4.3. Topology adapted from Liu *et al*. (2015), and Pfam alignment (Finn *et al.* 2016).

4.4.4. Flagellar and Membrane

Four isolates yielded variation within flagellar and membrane related genes, as detailed in Table 4.4.

Isolate D2-36 exhibited five instances of minor allele variation within locus CAMP0586. This locus – also referred to as Cj0628, yielded no alignments to the Pfam database. Cj0628 functions as a membrane bound autotransporter.

A single instance of minor allele variation was detected within the CAMP1178 locus, which encodes the major outer membrane protein, PorA.

Two isolates yielded variants within the CAMP1257 locus, which is responsible for flagellar glycosylation. Isolates D2-9 and D2-11 contained 4 and 5 instances of minor allele variation at this locus respectively. All substitutions in isolate D2-11 were transitional, whereas isolate D2-9 displayed two instances of synonymous substitutions. Residue 251 exhibited variation in both isolates, with both cases introducing a change to the amino acid sequence. All variants within this locus occurred at a relatively low frequency (0.14-0.28).

Table 4.4 - Minor allele variants detected within loci encoding flagellar and membrane related proteins

Isolate	Locus	Residue	Codon			Amino Acid			Mapping Quality*	Coverage Depth	MAF†
			Ref		Alt	Ref		Alt			
D2-36	CAMP0586	84	**T**TT	→	**G**TT	Phenylalanine	→	Valine	225	766	0.47
D2-36	CAMP0586	98	A**C**C	→	A**A**C	Threonine	→	Asparagine	225	822	0.47
D2-36	CAMP0586	98	AC**C**	→	AC**A**	Threonine	→	Threonine	225	815	0.46
D2-36	CAMP0586	99	**A**TA	→	**G**TA	Isoleucine	→	Valine	225	816	0.47
D2-36	CAMP0586	100	**A**AA	→	**G**AA	Lysine	→	Glutamic Acid	225	809	0.47
D4-7	CAMP1178	180	G**G**A	→	G**A**A	Glycine	→	Glutamic Acid	225	576	0.26
D2-9	CAMP1257	223	AG**C**	→	AG**T**	Serine	→	Serine	225	192	0.20
D2-9	CAMP1257	224	**C**AA	→	**G**AA	Asparagine	→	Glutamic Acid	225	190	0.20
D2-9	CAMP1257	229	TT**T**	→	TT**C**	Phenylalanine	→	Phenylalanine	225	190	0.20
D2-11	CAMP1257	251	**T**AT	→	**C**AT	Tyrosine	→	Histidine	182	142	0.14
D2-9	CAMP1257	251	T**A**T	→	T**G**T	Tyrosine	→	Cysteine	225	164	0.28
D2-11	CAMP1257	265	C**G**A	→	C**A**A	Serine	→	Asparagine	222	84	0.22
D2-11	CAMP1257	271	G**G**C	→	G**A**C	Glycine	→	Aspartic Acid	225	69	0.26
D2-11	CAMP1257	273	A**T**A	→	A**C**A	Isoleucine	→	Threonine	178	58	0.26
D2-11	CAMP1257	274	A**T**A	→	A**C**A	Isoleucine	→	Threonine	209	63	0.27

* - Provided as a PHRED score

† – Minor allele frequency (*i.e.* indicative of percentage of reads that support the minor allele).

Colours represent sequence type. Legend: ST-50 ST-44 ST-21

4.4.5. Homopolymeric Tract

Four isolates contained minor allele variants within the intergenic homopolymeric tract that mediates the CAMP1239 (Cj1321) locus, which encodes the protein Maf6. In all instances a C>G substitution occurred, converting a G10 tract to a G13 (Table 4.5), resulting in 'Off' expression.

Table 4.5 - Minor allele variation detected amongst multiple isolates in the Cj1321 homopolymeric tract

Isolate	Homopolymeric Tract Sequence			Mapping Quality*	Coverage Depth	MAF†
	Ref		Alt			
D2-32	GG**C**GGGGGGGGGG	→	GG**G**GGGGGGGGGG	35	111	0.2
D2-36	GG**C**GGGGGGGGGG	→	GG**G**GGGGGGGGGG	44	100	0.21
D4-24	GG**C**GGGGGGGGGG	→	GG**G**GGGGGGGGGG	73	67	0.28
D5-69	GG**C**GGGGGGGGGG	→	GG**G**GGGGGGGGGG	58	32	0.25

* - Provided as a PHRED score
† – Minor allele frequency (*i.e.* indicative of percentage of reads that support the minor allele).

4.4.6. Intergenic

Four instances of intergenic variation were detected, distributed across three isolates. All substitutions were transversions (Table 4.6).

Table 4.6 - Minor allele variation present at intergenic positions

Isolate	Nucleotide			Mapping Quality*	Coverage Depth	MAF†	Position
	Ref		Alt				
D2-36	**G**	→	**A**	205	819	0.45	Node_70 572
D3-22	**A**	→	**G**	208	347	0.43	Node_13 286
D3-22	**A**	→	**G**	225	372	0.53	Node_13 287
D5-80	**T**	→	**C**	118	258	0.34	Node_45 97

* - Provided as a PHRED score
†– Minor allele frequency (*i.e.* indicative of percentage of reads that support the minor allele).

4.4.7. Hypervariable Loci

The PubMLST *Campylobacter* database was interrogated for isolates containing whole-genome sequence data, and a confirmed presence of each variable loci identified in the deep sequencing dataset. A total of 56 isolates were returned and used for further analysis.

The dataset was broken down by reported allele number under the 1643 Gundogdu loci scheme (Gundogdu *et al.* 2007). Table 4.7 provides the ten most variable loci, as defined by the percentage of isolates exhibiting a unique allele.

Table 4.7 - Loci in which the greatest frequency of unique alleles was observed amongst a group of isolates from the PubMLST database

PubMLST Locus	Locus Aliases	Isolates With Unique Allele
CAMP1460	Cj1564, CcmL	71%
CAMP1257	Cj1341c, maf6	68%
CAMP1256	Cj1340c	63%
CAMP0035	Cj0036	50%
CAMP0038	Cj0039c, typA	50%
CAMP1178	Cj1259, porA	50%
CAMP0034	Cj0035c	48%
CAMP0124	Cj0136, infB	48%
CAMP0001	Cj001, dnaA	46%
CAMP0002	Cj002, dnaN	46%

4.4.8. Loci in relation to NCTC11168

In order to provide a common reference point, loci exhibiting minor allele variation were mapped against the common reference assembly of *C. jejuni* strain NCTC11168. The positions of these loci relative to NCTC11168 is provided in Figure 4.7.

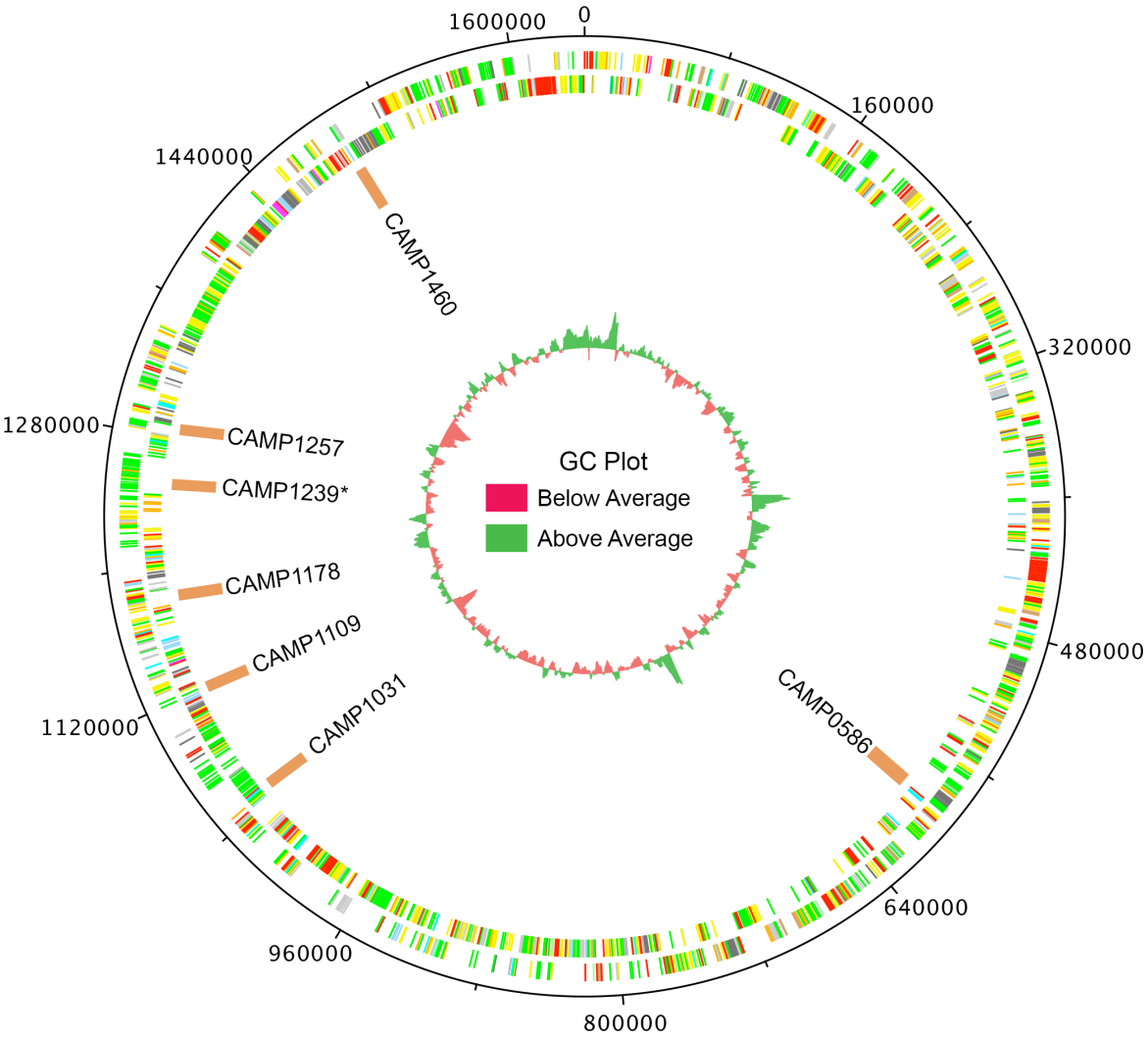

Figure 4.7 - Circular chromosome of NCTC111168 with annotated CDS*

CDS as annotated in the updated re-annotation of the complete NCTC111168 genome sequence (Gundogdu *et al.* 2007). Highlighted CDS were found to contain minor allele variation at one or more amino acid residues.

4.5. Discussion

The findings from this study demonstrate that single clinical populations of *Campylobacter jejuni* exhibit some degree of intra-populational diversity. A total of 17 isolates were analysed, yielding an average of ~2.1 polymorphisms per isolate. Whilst this is not as high as similar studies – for example studies of pathogens in Cystic Fibrosis patients (Lieberman *et al.* 2013, Smith *et al.* 2006, Workentine *et al.* 2013) – it is important to take the typical disease manifestation into account.

As discussed in the first chapter, the symptomatic phase of the disease lasts between 2 and 7 days (Blaser 1997). Even when considering extreme instances of incubation time, and the delay in which patients will typically report to clinicians, the length of time in which *Campylobacter* would have colonised the patient is in the order of weeks, compared to months and years in chronic conditions. Additionally, these isolates were prepared from a single stool sample. Patients pass between 8-10 stools per day (Blaser *et al.* 1983), meaning any one stool may conceivably not be completely reflective of the total diversity present within the host.

A previous study analysing the genetic heterogeneity of a human passaged *C. jejuni* NCTC11168 isolate only detected variation within homopolymeric tracts. Whilst this study also identified variation within one of these tracts, diversity was also apparent within several loci. This disparity in variation may be reflective of the clonality of the infecting population, or perhaps indicative of mechanisms by which variation is generated within the human host.

Many of the identified variants were found to occur at minor allele frequencies, supporting the diverse community model for intra-populational diversity (Lieberman *et al.* 2013). However, the methodology used within this study is not sensitive to fixed mutations. Previous use of the deep sequencing methodology involved mapping

genomic reads to a completed reference genome sequence (Lieberman *et al.* 2013).

In this study however, genomic reads were mapped against an assembly constructed using the same querying reads, and as such the existence of a fixed polymorphism is impossible to discern. Using a self-mapping approach is of great advantage when studying pathogens that cause acute pathologies, as sampling opportunities are limited by the clinical management of the disease. *Campylobacter spp*. also exhibit a large amount of genetic diversity, even between members of the same clonal lineage (Colles *et al.* 2003, Cody *et al.* 2013). This causes issues when selecting a sufficiently homologous reference sequence to query reads against. In chronic conditions, or in instances where repeat patient samples from a suitable time course are available, using a historic sequence of the same isolate as a reference would be preferable.

4.5.1. Chemotaxis Genes

The majority of identified variants were contained within genes related to chemotaxis. As discussed in Chapter 1, chemotaxis is an essential prerequisite for colonisation of the human gastrointestinal tract. Mutagenesis studies in mice, chickens, and human cell lines have all shown a reduced infection potential when key chemotaxis related genes are knocked-out (Rahman *et al.* 2014, Li *et al.* 2014, Hendrixson & DiRita 2004).

In *Campylobacter spp.,* chemotaxis related sensing is mediated by transducer-like proteins (Tlps) which bear a strong homology to the well documented Methyl-accepting chemotaxis proteins found in *E. coli* (Lertsethtakarn *et al.* 2011). These transducer-like proteins encode a variety of receptors that feed into the CheA-CheW-CheY signal transduction pathway, which ultimately leads to a change in the directional rotation of the flagellar motor (Korolik *et al.* 2008). A total of ten transducer-like proteins have been identified via genomic study (Marchant *et al.* 2002).

The high degree of genetic diversity present amongst chemotaxis related loci may be indicative of changes in selective pressure. As the disease progresses, fluid accumulates in the intestinal lumen and intestinal cells become damaged (Young *et al.* 2007, Wassenaar & Blaser 1999). This would likely alter the availability of certain nutrients and growth factors, as well as introduce an increased contact with immune-responsive cells.

4.5.1.1. CAMP1031

CAMP1031 – also referred to as Cj1110c – encodes transducer like protein 8 (Tlp8), CetZ. Tlp8 belongs (along with Tlp5 and Tlp6), to the group C transducer like proteins, which are unfortunately less well characterised than other Tlp groups (Korolik *et al.* 2008).

Predictive models of Tlp8 do not identify any transmembrane regions or signals of periplasmic export, which suggests this receptor is located in the cytoplasm (Reuter & van Vliet 2013). The CetZ protein contains two PAS-domains and a single methyl-accepting signalling domain. PAS-domains, named after the first genes in which these regions were identified (Per-Arnt-Sim), are responsible for sensing changes in cytosolic redox potential (Key *et al.* 2007).

This locus exhibited a single transversion from adenosine to guanine within amino acid residue 343. This leads to a non-conservative change in amino acid at this residue from the non-polar glycine to acidic residue glutamic acid. At this intragenic position, the most common allele represented in the PubMLST database (*i.e.* CAMP1031:1) expresses a glutamic acid residue. The substitution observed therefore may be a divergent haplotype that is either reverting to the most common allele, or a common haplotype undergoing variation in response to selective pressure.

This locus was also shown to undergo polymorphic change in other studies. Genomic investigation of a milk-borne outbreak amongst six patients revealed numerous polymorphic loci, with CAMP1031 displaying

allelic diversity between the original source and subsequent patient isolates (Revez *et al.* 2014). Similarly, in accidental repeated sampling of 17 patients, Cody *et al.* (2013), reported allelic variation in two patients. Comparison in these studies was performed using wgMLST, and as such the exact genotypic change and position is unknown. However, as wgMLST is assembly based, it suggests that the disparate alleles were present at a majority frequency that permitted inclusion in the final assembly. There may in fact have been a far greater diversity present in these samples at minor allele frequencies which would be important to consider in future studies.

Variation within this locus may occur as the result of changing metabolic health within *Campylobacter* cells. The internal physiological state of the cell may be a driving factor in the modulation of other chemotaxis pathways. Reuter *et al.* (2013), propose that CetZ is responsible for redox sensing, and may interact with a redox metabolite such as FAD, which is seen in other proteins containing PAS domains (Key *et al.* 2007). The polymorphic base was located within the methyl-accepting signalling (MA) domain, which is responsible for signal transduction in the chemotaxis pathway. The lack of variation within the PAS domain indicates that the ligand recognition site is conserved. Variation within the MA-signalling domain suggests that a haplotype is present in the population that may exhibit modulated transduction of the ligand-binding signal. This could for example alter the priority of the signal transduction pathway and alter ligand-specific motor switching by increasing, deleting, or reducing the signal transduction potential of the CAMP1031 locus.

The polymorphism detected within this locus falsely appears to occur in the majority of reads (MAF 0.58), however it's important to note that the way in which MAF is calculated only considers bases sequenced with a Q-score greater than 30 (*i.e.* 99.99% accuracy). Reads were processed using a sliding window quality trimmer, which was configured to trim 25-bp regions that fell below an average Q score of 20 (*i.e.* 99.9% accurate), and whilst the mapping process further qualifies bases at Q30, the assembler

does not. It is likely that the two divergent alleles are present at an intermediate frequency, leading to ambiguity in the final assembly. This highlights the limitations of assembly based comparisons – in the absence of the allele frequency, it would appear that this variant was fixed amongst the population, however in reality these alleles are coexistent.

4.5.1.2. CAMP1109

CAMP1109 encodes Tlp9, and is the lone member of group B transducer like proteins. It is also referred to as CetA, for *Campylobacter* energy taxis protein A (Korolik *et al.* 2008). The structure of CetA suggests it is a cytoplasmic protein, and is anchored to the membrane via a trans-membranous N-terminal region (Marchant *et al.* 2002, Korolik *et al.* 2008). A single minor allele variant was detected within this locus, with residue 275 exhibiting a conservative mutation between the polar residues alanine and valine.

The *cet*A locus is essential for efficient energy taxis, with knock out studies revealing reduced taxis towards sodium pyruvate and fumarate (Reuter & van Vliet 2013). The overall mechanisms controlling energy taxis are also mediated by CetA. Research suggests that efficient energy taxis occurs due to a balance of activity between the *cet*A, *cet*B, *cet*C and *cet*Z loci, with the combination of signals from these genes conferring appropriate taxis towards favourable energy generation conditions and away from hazardous redox potentials (Reuter & van Vliet 2013). Both the *cet*A and *cet*Z loci yielded variation within the same isolate, which would support the hypothesis presented by Reuter and Van Vliet (2013).

4.5.1.3. CAMP1460

The locus CAMP1460 – also referred to as Cj1564 or *ccmL* – yielded minor allele variation across multiple ST-21 isolates. This locus encodes Tlp3, a group A chemoreceptor responsible for ligand binding in the chemotactic sensory pathway. Group A Tlps share a structural homology with the methyl-accepting proteins found in *E. coli* and contain both transmembrane and periplasmic domains, suggesting that they may

sense ligands that are external to the cell (Marchant *et al.* 2002). Tlp3 is also known as ccmL or *Campylobacter* chemoreceptor for multiple ligands, and interacts with several chemoattractants (*e.g.* isoleucine, purine, malic acid) and chemorepellents (*e.g.* lysine, arginine, glucosamine). *In vitro* analysis demonstrates that the receptor is able to bind multiple ligands simultaneously (Rahman *et al.* 2014).

The positions at which variable amino acid residues occur reside within the intracellular MA-signal transduction domain, as observed in the other chemotaxis loci. Again, this may result in modulation of the strength of a chemotaxis response to a particular ligand. For example, a phenotype that has a reduced ability to transduce chemotactic signals for a ligand that is found in abundance may lead to other loci taking priority over the flagellar switching mechanism. This in turn would lead to a greater chemotaxis towards less common ligands, and a modulated growth potential.

Two amino acid residues underwent repeat variation in separate isolates. Residue 523 exhibited separate base substitutions in the first codon position, resulting in a non-conservative change from threonine to non-polar residues alanine and proline. Similarly, at residue 550 the second codon position underwent an identical base change, resulting in a conservative substitution between polar residues leucine and proline. Repeat variation – particularly with the same amino acid substitution – in independent isolates suggests that this locus may be encountering (and adapting) to a specific selective pressure.

The frequency of alleles at these common residues varies, with isolate D4-69 yielding a MAF of ~0.3, vs ~0.5 in isolate D2-69. This may be indicative a beneficial mutation sweeping to fixation, with sampling of the two isolates occurring at different stages of the disease progression. The CAMP1460 locus appears to be highly variable in general - it was the most variable locus identified amongst WGS isolates submitted to the PubMLST *Campylobacter* database, with 71% of isolates containing a unique allele – this supports the notion that variation within this locus

typically moves to fixation. Additional sampling of a single patient at multiple time points would aid in delineating whether this is the case, or whether perhaps these alleles would continue to exist at an intermediate frequency under the diverse community model of variation (Lieberman *et al.* 2013). Given that the frequency of several alleles detected within this locus are actually present in the majority of high-quality reads, the diversity observed in the PubMLST may be a consequence of the *de novo* assembly process, as discussed in regards to CAMP1031.

In a mutagenesis study, knock out of the *CcmL* locus created a phenotype with altered cellular shape and reduced chemotactic ability. The *CcmL* mutant also displayed a greater affinity for autoagglutination, and increased biofilm formation (Rahman *et al.* 2014). *Campylobacter* is pleomorphic, and is known to lose its helical shape in times of metabolic stress. This conformational change is also associated with an increase in biofilm formation (Ikeda & Karlyshev 2012, Walker *et al.* 1986). It's possible that the variation observed within the *ccmL* locus had some role in shifting the conformational state of the bacteria.

4.5.2. Membrane and Flagella

Three loci were found to contain minor alleles with functions pertaining to the membrane or flagella: CAMP1178, CAMP0586 and CAMP1257.

4.5.2.1. CAMP1178

The CAMP1178 locus encodes the major outer membrane protein (MOMP), PorA, which regulates membrane permeability to small molecules (Cody *et al.* 2009). Porins are composed of several α-strands which form β -barrels. The bottom of these β-barrels exhibit short turns which face towards the periplasmic space, and at their top side express long loops that are external to the cell surface membrane (Stathopoulos 1999).

Variation was observed at amino acid residue 180. This residue underwent a non-conservative substitution of glycine with glutamic acid

as result of a G>A transition in the second codon position. Residue 180 is located within the putative surface exposed loop 7 (Cody *et al.* 2009, Zhang *et al.* 2000). Research shows that the *por*A locus undergoes variation during prolonged infection with *Campylobacter jejuni* (Cody *et al.* 2009). A single patient provided stool samples that were collected 16 days apart, with NGS analysis revealing a fixed variation at residue 170. This residue also exhibited a G to A transition in the second codon position, causing glycine substitution with aspartic acid.

Both the variable residues in this study, and the work from Cody *et al.* (2009), were located within the surface exposed loop 7, and in both instances variation induced a G>A transition at the second position of a glycine codon. Glycine is a hydrophobic aliphatic residue, and is demonstrated to be important in the packing of α-helices (Senes *et al.* 2000). In both instances of variation, glycine was replaced with the negatively charged residues glutamate and aspartate, which may lead to conformational change to the loop structure. The immune response to *Campylobacter spp.* is largely reliant on the recognition of surface-exposed conformational epitopes (Huang *et al.* 2007). The variation within the surface exposed loops may be a result of immune exposure, with conformational change potentially modulating immune recognition of *Campylobacter* cells.

The polymorphic base in this study was detected at a frequency of 0.26 – in the prolonged infection a similar variant was found to sweep to fixation. This may suggest that sustained exposure to host-immunity exerts purifying selective pressure, driving antigenic variation within the loop 7 structure. It would be interesting to apply the deep sequencing methodology in a case of prolonged exposure, with samples collected at multiple points during the symptomatic phase of the disease. This would allow tracking of the frequency at which any variation occurs, and delineate variants undergoing positive and negative selection by revealing alleles that move from minor to major frequencies.

4.5.2.2. CAMP0586

The CAMP0586 locus (also referred to as Cj0628), encodes an autotransporter protein known as CapA. In Gram-negative bacteria, autotransporter proteins are ubiquitous, and are responsible for several virulence mechanisms, including adhesion and toxigenicity (Henderson *et al.* 2004). CapA is a surface exposed lipoprotein, located on the outer cellular membrane (Ashgar *et al.* 2007). The CapA lipoprotein has been shown to mediate infection of human epithelial cells, with knock-out mutants expressing decreased adhesion in Caco-2 cell lines (Ashgar *et al.* 2007). A similar result was observed in chicken colonization, with *cap*A mutants exhibiting a 75% reduction in whole-flock positivity (Ashgar *et al.* 2007).

Five instances of polymorphic alleles were detected in a single deep-sequenced isolate. One of these variants at residue position 98 resulted in a synonymous substitution, with the remaining variants causing non-synonymous mutations. The 98th residue yielded a second polymorphic base with a similar allele frequency to the first (0.47 and 0.46 respectively). Considered individually, this would have led to a conservative substitution of polar residues threonine to asparagine. The intermediate minor allele frequency indicates that there may be two concurrent lineages in circulation, and as such it is also possible that both bases were polymorphic as part of a single haplotype. This would lead to the codon of residue 98 shifting from ACC to AAA, causing a non-conservative substitution of threonine to lysine. This, along with the non-conservative substitution of lysine with glutamate at residue 100 may have produced a conformational change.

This locus is the second surface-exposed membrane protein exhibiting variation, which as discussed regarding *por*A, may be a result of exposure to host immunity. PorA has been shown to contain multiple putative epitopes based on protein sequence analysis (Ingale & Goto 2013), which suggests this protein is of antigenic importance. The variation within this locus may again be an example of epitope polymorphism.

4.5.2.3. CAMP1257

CAMP1257 – also referred to as Cj1341c – encodes an unknown protein referred to as motility accessory factor 6 (maf6). Maf6 is part of the *maf* gene family, and is located adjacent to the *fla*A flagellin gene. Other *maf* family members modulate flagellin variation via glycosylation, or slip strand mispairing as a result of repeat homopolymeric tracts, however the function of maf6 has not been elucidated (Wassenaar *et al.* 1995, Szymanski *et al.* 2003).

This locus exhibited variation across two isolates, with a single residue (251) undergoing repeat variation in both. All variants were in a 150bp region of CDS (covering amino acid residues 223 to 274), though no conserved domains were identified via Pfam alignment. The variation present in this locus, and the paucity of data on its function makes it an interesting target for future study.

4.5.3. Homopolymeric Tracts

Campylobacter spp., as well as other organisms such as *Neisseria,* are shown to mediate the expression of phase variable genes via the alteration of repetitive DNA sequences (J. Parkhill *et al.* 2000). These tranches of repeat bases are referred to as homopolymeric tracts, and induce slip strand mispairing during DNA replication, which ultimately leads to a shift in reading frame and differential expression of the affected gene (Henderson *et al.* 1999). The 11168 *Campylobacter* genome for example, contains 28 phase variable genes that are mediated by a range of poly-G tract lengths (Lango-Scholey *et al.* 2016).

The deep sequencing dataset revealed four isolates exhibiting minor allele variation within a single homopolymeric tract, that mediates expression of the CAMP1239 (Cj1321) locus. This tract is one of three intergenic poly-G tracts that have been identified as mediating the expression of phase-variable genes within *Campylobacter spp.* (Aidley & Bayliss 2014).

The Cj1321 homopolymeric tract confers 'on' expression when the G-tract length covers 10 base-pairs. In all four instances, a minor allele variant of the poly-G tract identified a G-10 to G-13, on-to-off alteration. This change in expression was conferred via a C > G substitution of the same base pair in positional relation to the G-10 tract.

The Cj1321 locus encodes an acetyltransferase which is located within the O-linked flagellar glycosylation locus, and has been shown to modulate the flagellin protein via pseudaminic acid modification (Champion 2005, Kim *et al.* 2012). The differential expression of the Cj1321 locus conferred by the minor allele may be related to biofilm formation. In mutagenesis studies, knock out of the O-linked flagellar glycosylation locus resulted in decreased biofilm formation and chick colonisation, however it did not affect motility (Howard *et al.* 2009).

Homopolymeric tracts are found to be hypervariable, and are often identified as undergoing variation in a variety of sources (J Parkhill *et al.* 2000, Bayliss *et al.* 2012, Revez *et al.* 2013, Cody *et al.* 2013, Cody *et al.* 2009). It is perhaps strange that only a single tract was found to undergo variation, but this may be explained by limitations of the deep-sequencing methodology. Whilst still above the Q30 quality threshold (*i.e.* 99.99% accuracy), the mapping quality of the Cj1321 tract variants was lower than variants discovered in other loci. This is likely as a result of decreased statistical confidence in stretches of single base repeat sequence. Additional instances of minor allele variation may have occurred within other poly-N tracts, but did not present with sufficient quality to be detected by the deep-sequence pipeline.

4.5.4. Future work

This study provides novel insights into the intra-population diversity of clinical *Campylobacter* infection. However, the study only includes a single CC – ST-21. As the most commonly represented lineage in human infection, it is unclear as to whether these isolates would undergo less variation – as they may already have a level of adaptation to human passage – or whether they are successful due to their ability to generate variation. To investigate this, expanding the dataset to include other sequence types that are less common would provide a more comprehensive overview of *Campylobacter* diversity.

Additionally, this study does not investigate whether the observed diversity is present amongst the population *before* clinical infection. For example, the degree of variation reported may be sustained from the chicken host through to consumption and propagation of human disease. Applying the deep sequencing methodology to an infectious population before and after human passage would aid in delineating the origin of diversity.

Chapter Five: Genetic heterogeneity of *Campylobacter spp.*

5.1. Abstract

In the previous chapter, clinical isolates from the dominant clonal lineage of *C. jejuni (*ST-21) were demonstrated to exhibit intra-population diversity at a number of key loci. This presented questions as to whether diversity could also be detected in other CC's, or whether it was specific to - and perhaps responsible for the abundance of - ST-21. It was also unclear as to whether the diversity observed was generated during the course of human disease, or perhaps maintained from a genetically diverse infectious population.

To address these issues, this chapter expands on the ST-21 dataset by sequencing clinical isolates from a further five CC's, including the *C. coli* lineage, ST-828. Additionally, we provide an insight into the diversity of *Campylobacter* populations prior to human infection, by sequencing populations obtained from the largest single source of human campylobacteriosis: fresh retail chicken.

Sequencing of the additional CC's revealed several loci undergoing minor allele variation that are exclusive to non ST-21 lineages. The major functional group of variable loci in both ST-21 and non ST-21 lineages was comprised of chemotaxis genes. Furthermore, two chemotaxis loci encoding transducer like proteins (Tlp) 3 and 8 yielded minor allele variation in both ST-21 and non ST-21 complexes. Tlp3 in particular contained minor allele variants in ST-21, ST-48, and ST-828 isolates, making it the only locus to be repeatedly identified in both multiple lineages and in multiple species. The overall dominance of the chemotaxis loci amongst the clinical datasets suggests that they are coming under strong selective pressure within the gastrointestinal environment, and as shown in Tlp3, may exhibit variation that is exclusive to the human host.

In contrast to the clinical datasets, the poultry isolates did not contain any variation within chemotaxis loci. Less variation was detected overall, and loci that were found to contain variation were linked to membrane and flagellar functions. This likely reflects the harsh ecological niche of retail poultry meat, and may be a result of significant selective bottlenecks during the slaughter and retail environments.

These results suggest that poultry related campylobacteriosis likely occurs via relatively homogenous populations, and that the diversity observed amongst clinical isolates may in fact be generated during or in response to human infection. Additional sequencing of non-clinical *Campylobacter spp*. would be an important direction for future work to determine the validity of this hypothesis.

5.2. Introduction

Clinical cases of campylobacteriosis are largely dominated by the ST-21 complex of *C. jejuni* (Cody *et al*. 2012; Kovanen *et al*. 2014; Schielke *et al*. 2014; Sheppard *et al*. 2009). In Chapter 3, we found that this was also true of isolates obtained from the greater Nottinghamshire area, with ST-21 forming the largest constituent complex.

In Chapter 4, we demonstrated that clinical ST-21 isolates were composed of genotypically diverse populations. This represents the first description of intra-population diversity amongst routine clinical isolates, and identified a number of loci relating to chemotaxis, membrane and flagellar proteins that were found to contain minor allele variation. Due to the novelty of this finding, there is no comparable data from isolates belonging to other CC's, and as such it is unclear as to whether the observed diversity is unique to ST-21 isolates.

Additionally, there is no comparable data for *Campylobacter coli* isolates. *Campylobacter coli* is responsible for approximately 7-12% of all clinical campylobacteriosis (Chen *et al*. 2011, Gillespie *et al*. 2002, Olson *et al*. 2008), which represents a significant number of illnesses. The *C. coli* CC ST-828 was found to be the second most common amongst the clinical dataset presented in Chapter 3, and is found in similar abundance in other epidemiological datasets (Cody *et al*. 2012).

In order to address these issues, we sought to apply the deep sequencing methodology to additional isolates from non ST-21 clonal lineages. Samples were selected from the genetic epidemiology work presented in Chapter 3, with representative isolates from several CC's found to occur in more than 5% of the dataset. This included several isolates belonging to the *C. coli* ST-828 complex.

5.2.1. Campylobacteriosis and Poultry

Poultry meat is the greatest single source of campylobacteriosis, largely arising from the high prevalence of retail packages found to be contaminated with *Campylobacter spp.*, and the significant volumes of meat consumed by the British public (Wilson *et al.* 2008, Corry & Atabay 2001, Key Note 2015). Chicken meat accounts for approximately 70-80% of poultry production, and the chicken reservoir alone is thought to account for between 50-80% of human *Campylobacter spp.* infections (Skarp *et al.* 2016).

Poultry manufacturers go to extensive lengths to try and reduce the level of *Campylobacter spp.* contamination present on the final retail product. At the farm level, several biosecurity measures (*e.g.* coop decontamination, worker hygiene) are employed with varying degrees of efficacy (Newell *et al.* 2011, Gibbens *et al.* 2001). However, despite these efforts *Campylobacter spp.* is found to routinely colonise approximately 75-90% of British commercial flocks (Humphrey *et al.* 1993, Evans & Sayers 2000, Stern *et al.* 2001). This contamination persists during the slaughtering process, which ultimately results in approximately 71% of retail chicken found to be colonised with *Campylobacter spp.* (Food Standards Agency 2016).

In the previous chapter, clinical isolates of the dominant clonal lineage - ST-21 - exhibited several instances of intra-population diversity, which provided the first evidence of genetic heterogeneity amongst routine clinical cases. However, it is not clear as to whether this heterogeneity is generated during the course of human infection, or simply maintained from a genetically variable antecedent population. In order to address this, the deep sequencing methodology would need to be applied to the original aetiological isolate prior to human infection, to allow an appropriate comparison. However, given the sporadic nature of clinical campylobacteriosis, source attribution and therefore source sampling presents a significant challenge.

In the absence of a suitable ancestral isolate, we sought to apply the deep sequencing methodology to *Campylobacter* populations that would typically be encountered by the public, and logically may lead to infection. Given that poultry – and specifically chicken meat - is the largest single source of human campylobacteriosis, we collected *Campylobacter spp*. isolates from samples of fresh retail chicken to provide a comparison between a potentially pre-infectious source, and the post-infectious clinical datasets.

5.3. Methods

5.3.1. Poultry

Additional isolates were obtained from fresh retail poultry purchased from large-chain retailers. As chicken is widely available and the most common poultry meat consumed in the UK, samples were restricted to chicken only. Retail chicken sold in the United Kingdom is labelled with a mandatory 'approved premises' identifier code, which provides information on the supplier of the poultry meat. In order to ensure that no single batch of poultry meat (and potentially the same microbial population) was repeatedly sampled, this code along with the expiry date were used to exclude packages that may have originated from a single batch. A variety of chicken parts were sampled, including breast, thigh and leg meat both with skin intact and removed. All meat was transported to the laboratory in under 30 minutes, and stored at 4°C for future use.

The chicken meat was weighed, placed inside a sterile stomacher bag, and submerged in maximum recovery diluent (Oxoid, United Kingdom, CM0733) at a ratio of 1ml:1g. The surface of the meat was manually massaged through the stomacher bag for two minutes to increase the amount of bacteria shed from the meat surface. The diluent was then decanted into sterilised conical flasks, and placed on a shaking platform at 100RPM for 1 minute. This stock solution was diluted in serial from 10^{-1} to 10^{-5} using fresh maximum recovery diluent.

These dilutions were then plated in replicate on to *Campylobacter* Blood-Free Selective Agar Base (Oxoid, United Kingdom, CM0739), supplemented with 16mg/500ml of Cefoperazone and 5mg/500ml of Amphotericin B (Oxoid, United Kingdom, SR0115). These agar plates were prepared according to the methodology previously described in Chapter 2. Isolates were incubated at 37°C for 48 hours in a microaerobic workstation (Don Whitley Scientific, United Kingdom, DG250). *Campylobacter* colonies were selected for deep sequencing analysis based on colony morphology and oxidase status.

5.3.2. Clinical Isolates

Human clinical isolates belonging to non ST-21 CC's – including four isolates from the *C. coli* complex ST-828 - were selected randomly from the genomic dataset produced in Chapter 3. These isolates were then prepared for subsequent deep sequence analysis.

5.3.3. Deep Sequencing

As a brief reminder, the deep sequencing methodology expands on conventional WGS analyses (*e.g.* wgMLST, SNP-calling) by greatly increasing the sequencing coverage depth (*i.e.* number of times a kmer is sequenced). This additional coverage confers a greater resolution of minor frequency variants that are contained within genetically heterogeneous sample populations. The deep sequencing methodology, as well as details concerning isolate culture, DNA extraction, WGS, and bioinformatics analyses are discussed in greater depth in Chapters 2, 3 and 4.

5.4. Results

5.4.1. Non ST-21 Isolates

5.4.1.1. Isolates, sequencing and assembly.

A total of 11 non ST-21 isolates were sequenced to sufficient quality and coverage, with an average depth of 241X. All isolates were found to be obtained from separate patients by age, sex (Table 5.1) and post code (data not shown).

Table 5.1 Patient demographics of samples belonging to the non ST-21 deep sequencing dataset

Isolate	CC	Collection Month	Patient Sex	Patient Age	Coverage Depth
D1-29	48	January	F	25	281
D1-35	464	January	F	67	120
D3-40	48	April	M	9	281
D4-57	45	June	F	78	177
D6-33	353	August	M	24	225
D7-17	828	August	M	43	305
D7-24	828	August	M	33	388
D7-31	828	September	M	55	182
D7-40	828	September	F	22	197
D8-67	353	October	F	63	284
D9-60	48	November	M	84	215

Samples were screened to determine whether isolates were obtained from separate patients. In instances where this was not clear by the provided demographics, geographic data was used for exclusion (data not shown).

The non ST-21 deep sequencing dataset had an average N50 of 309,984bp (95% C.I: 242,311bp to 377,657bp), a mean total length of 1,816,875bp (95% C.I: 1,691,046bp to 1,942,705bp) and a mean number of contigs of 177.2 (95% C.I: 5.872 to 348.5). Figure 5.1 shows the distribution of these quality metrics.

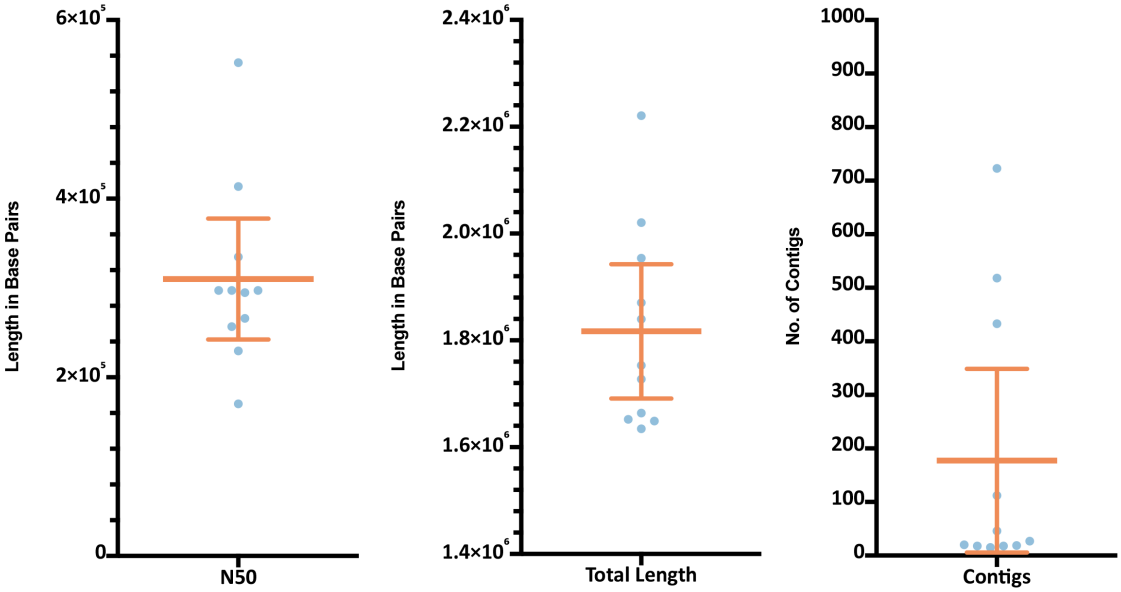

Figure 5.1 - Quality metrics of draft *de novo* assemblies of non ST-21 isolates

Mean inclusive of upper and lower confidence intervals highlighted in orange.

Linear regression was used to assess whether the sequencing coverage depth, assembly N50 or number of contigs had an impact on the number of loci in which minor allele variation was detected (Figure 5.2). These analyses revealed no statistical significance (P = 0.6993, 0.9923 and 0.2322 respectively).

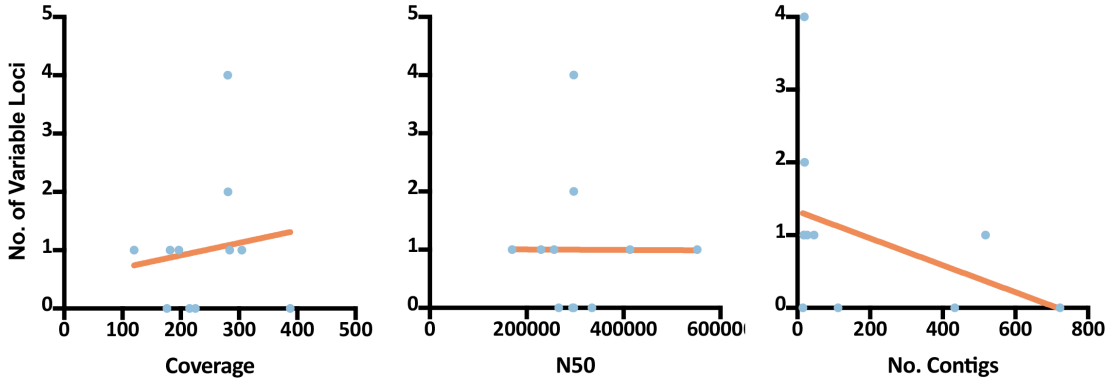

Figure 5.2 - The statistical impact of sequencing and assembly metrics on the number of loci in which minor allele variation was detected

*Statistical significance calculated using a linear regression analysis, in all instances results were statistically insignificant (Coverage: P = 0.6933, N50: P = 0.9923, No. Contigs: P = 0.2322)

5.4.1.2. Minor allele variation

The non ST-21 deep sequencing dataset contained 20 minor allele variants with sufficient depth and quality to qualify further analysis. Of the 11 isolates sequenced, 36% (N=4) did not yield any minor allele variants that passed the quality filtering process

18 of the 20 variants were contained within identified loci. The loci in which minor allele variation was detected are detailed below in Table 5.2.

Table 5.2 - Genes in which minor allele variation was detected within non ST-21 dataset, grouped according to putative function

Isolate	Locus	Functional Group	Residue	Codon Ref		Codon Alt	Amino Acid Ref		Amino Acid Alt	Mapping Quality*	Coverage Depth	MAF†
D1-29	CAMP0252	Chemotaxis	34	**G**GC	→	**A**GC	Glycine	→	Serine	66	859	0.17
D7-40	CAMP0252	Chemotaxis	292	**A**TG	→	**G**TG	Methionine	→	Valine	112	386	0.22
D1-29	CAMP0410	Chemotaxis	172	**G**CC	→	**A**CC	Alanine	→	Threonine	35	537	0.18
D1-29	CAMP1031	Chemotaxis	358	**G**CT	→	**A**CT	Alanine	→	Threonine	90	484	0.2
D3-40	CAMP1460	Chemotaxis	475	AA**C**	→	AA**T**	Asparagine	→	Asparagine	225	416	0.39
D7-17	CAMP1460	Chemotaxis	482	TC**T**	→	TC**A**	Serine	→	Serine	113	2891	0.27
D7-17	CAMP1460	Chemotaxis	502	AC**C**	→	AC**T**	Threonine	→	Threonine	208	3241	0.27
D7-17	CAMP1460	Chemotaxis	534	AA**A**	→	AA**G**	Lysine	→	Lysine	91	3282	0.26
D7-17	CAMP1460	Chemotaxis	606	**A**GT	→	**T**GT	Serine	→	Cysteine	225	2781	0.48
D7-17	CAMP1460	Chemotaxis	606	A**G**T	→	A**C**T	Serine	→	Threonine	225	2794	0.48
D7-17	CAMP1460	Chemotaxis	609	GA**A**	→	GA**G**	Glutamic acid	→	Glutamic acid	225	2548	0.48
D7-31	CAMP0049	Membrane	18	AG**C**	→	AG**T**	Serine	→	Serine	225	159	0.64
D7-31	CAMP0049	Membrane	20	GA**A**	→	GA**G**	Glutamic acid	→	Glutamic acid	225	171	0.66
D3-40	CAMP0922	Membrane	35	**G**TT	→	**A**TT	Valine	→	Isoleucine	129	139	0.2
D8-67	CAMP1055	Membrane	121	AC**A**	→	AC**C**	Threonine	→	Threonine	225	293	0.5
D1-29	CAMP0303	Flagellar	190	G**T**T	→	G**C**T	Valine	→	Alanine	177	753	0.29
D1-35	CAMP1213	Flagellar	279	CT**A**	→	CT**G**	Leucine	→	Leucine	99	110	0.28
D1-35	CAMP1213	Flagellar	281	G**A**T	→	G**G**T	Aspartic acid	→	Glycine	108	114	0.3

* - PHRED-quality scores. † - Minor allele frequency, *i.e.* proportion of reads supporting the alternative allele. Colours represent clonal complex.
Legend: ST-48 ST-828 ST-353 ST-464.

5.4.1.2.1. Chemotaxis Loci

Three isolates yielded variation within four different chemotaxis related genes: CAMP0252, CAMP0410, CAMP1031 and CAMP1460. The two latter loci were previously found to undergo variation in the previous deep sequencing dataset (Section 4.4), which applied the same methodology to the most common clinically observed CC, ST-21.

Two separate isolates yielded minor allele variation within the CAMP0252 locus (Table 5.2). This locus – also referred to as Cj0283c – encodes the scaffold protein, CheW. Isolate D1-29 varied at amino acid residue 34, which is contained with the Hpt domain. Isolate D7-40 was found to be variable at residue 292, which is located with the signal transducing histidine kinase domain (Figure 5.3).

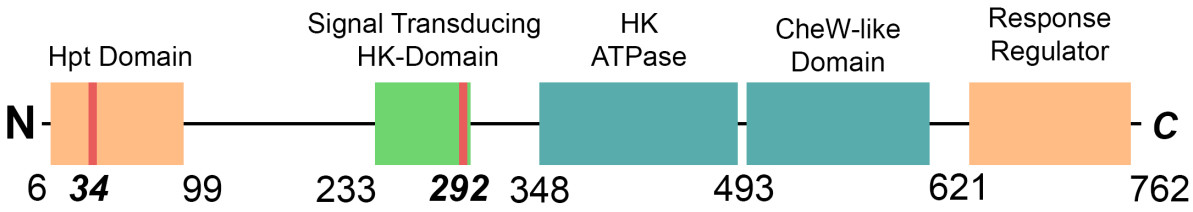

Figure 5.3 - Topology of the CAMP0252/Cj0284c locus

Residue numbers provided below domains show regions of Pfam alignment. Residue 34 and 292 yielded variation at the first base position, shown in Table 5.2.
Hpt= Histidine Kinase Phosphotransfer
HK = Histidine Kinase
ATP = Adenosine Triphosphate

CAMP0410 – also referred to as Cj0448c - encodes the group C transducer like protein Tlp6 (Marchant, Wren, & Ketley, 2002). Tlp6 is one of the less well characterised chemotaxis proteins, however in general it is believed that group C Tlps are responsible for detecting cytoplasmic signals (Lertsethtakarn, Ottemann, & Hendrixson, 2011; Zautner, Malik Tareen, Groß, & Lugert, 2012).

A single isolate (D1-29) was found to contain a minor allele variant at amino acid residue 172. Protein sequence alignment to the Pfam database revealed that this variant was contained within a conserved methyl accepting signalling domain.

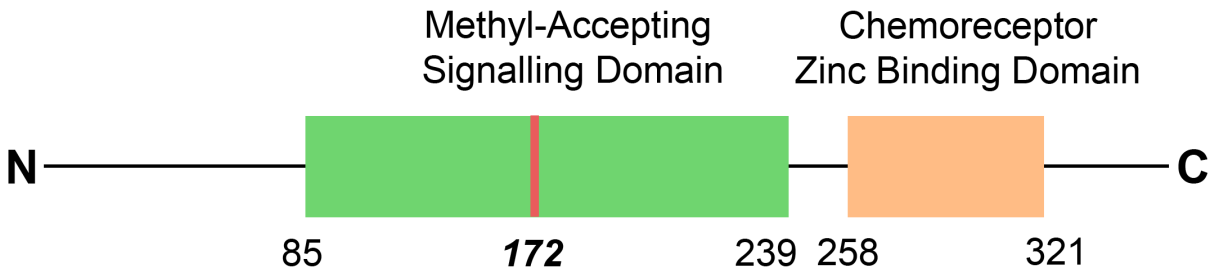

Figure 5.4 - Topology of the CAMP0410/Cj0448c locus

Residue numbers provided below domains show regions of Pfam alignment. Residue 172 yielded variation at the base position, shown in Table 5.2.

CAMP1031 – or Cj1110c – encodes Tlp8 (CetZ). Residue number 358 contained a minor frequency variant in the first codon position, causing a non-synonymous mutation from alanine to threonine. This residue is located within the methyl-accepting signalling domain of the protein. This locus was also observed to undergo variation at residue 343 in ST-21 isolate D5-69, represented by the blue demarcation in Figure 5.5.

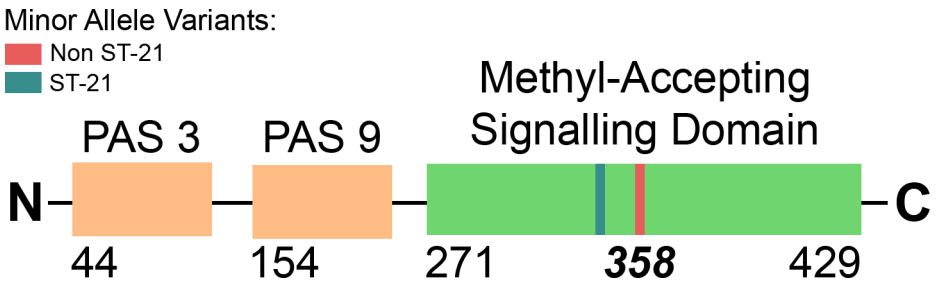

Figure 5.5 - Topology of Tlp8 (CetZ) encoded by the CAMP1031/Cj1110c locus with SNP data from ST-21 and Non ST-21 isolates.

Residue numbers provided below domains show regions of Pfam alignment. Residue 358 yielded variation at its first base position, shown in Table 5.2.This locus also contained a minor allele variation in Chapter 4's ST-21 dataset, with the variant residue shown in blue. Per-ARNT-Sim (PAS).

Two non ST-21 isolates yielded minor allele variants within the CAMP1460 locus. CAMP1460 – also referred to as Cj1564 or *ccmL* – encodes transducer like protein 3 (Tlp3). All minor allele variants were contained within methyl-accepting signalling domain, as shown in Figure 5.6. Two variants occurred at the same amino acid residue (606), and at the same minor allele frequency (0.48) in isolate D7-17.

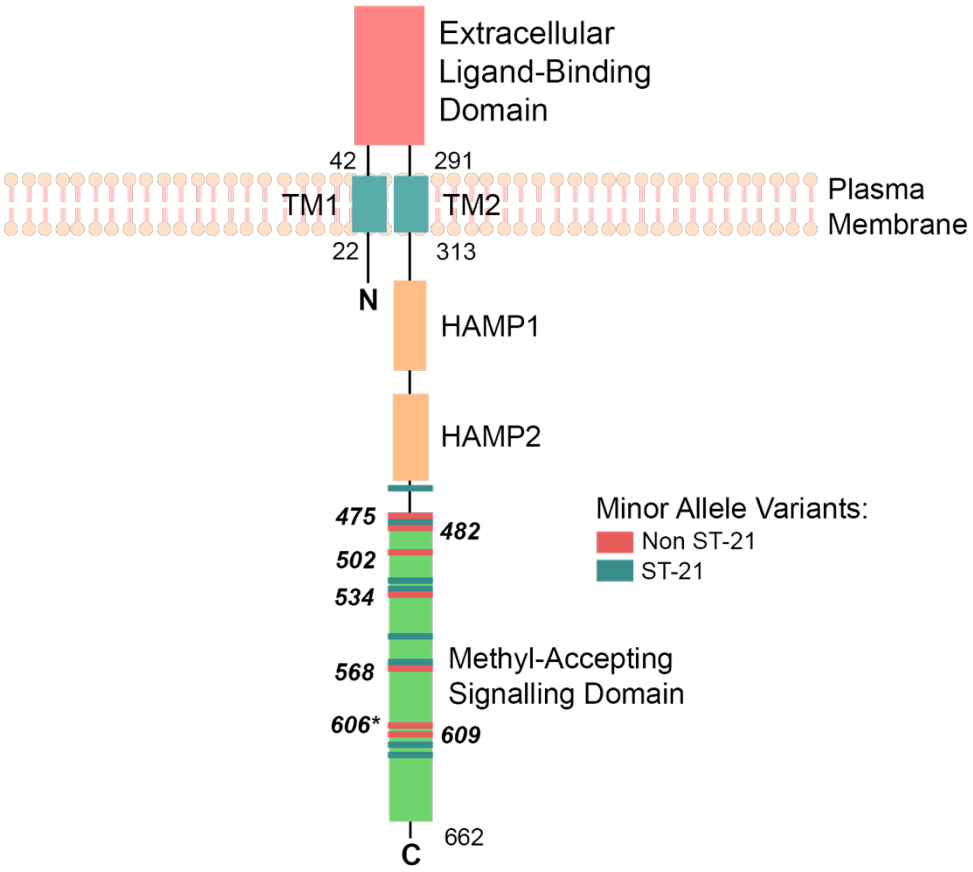

Figure 5.6 - Topology of Transducer-like protein 3 (CcmL) highlighted with positions of variable residues as provided in Table 5.2

* - Residue 606 identified as variable at the first two codon positions.
TM1 and 2 = **t**rans**m**embrane helixes.
HAMP1 and 2 = **h**istidine kinases, **a**denylyl cyclases, **M**CPs and **p**hosphatases.
Residue positions provided in bold italicised font adjacent to red lines correspond to minor allele variations detected and listed in Table 5.2.
Residues highlighted in blue provide approximate locations of variable residues identified in Chapter 4's ST-21 dataset.
Topology adapted from Liu *et al*., 2015 and Pfam alignment (Finn *et al.* 2016).

5.4.1.2.2. Membrane Loci

Three isolates were found to contain minor allele variation within loci that have gene products related to the cellular membrane.

CAMP0049 (Cj0057) encodes a putative periplasmic protein. Isolate D7-31 contained two minor allele variants within this locus at positions 18 and 20 (Table 5.2). Protein alignment to the Pfam database did not identify any conserved domains.

CAMP0922 – also referred to as Cj0999c – encodes a putative integral membrane protein. Protein alignment revealed a single conserved domain – 698 – in which a minor allele variant was discovered (Figure 5.7). This domain has no resolved function, however it belongs to a group of proteins involved in transport (CL0064). CAMP0922 has yet to be characterised, and no studies have been performed on the locus to date

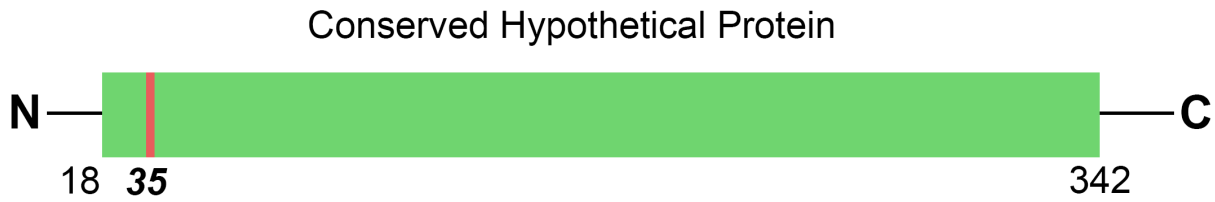

Figure 5.7 - Topology of the CAMP0922/Cj0999c locus encoding a putative integral membrane protein

Residue numbers provided below domains show regions of Pfam alignment. Residue 35 yielded variation at first base position (Table 5.2) contained within a conserved domain linked to membrane transport.

CAMP1055 – also referred to as Cj1134 or *htr*B - encodes a putative acyltransferase involved in lipid A biosynthesis. This locus contained a single synonymous mutation at the 121st amino acid residue, which is located within the conserved Acyltransferase domain that spans the majority of the CDS (Figure 5.8).

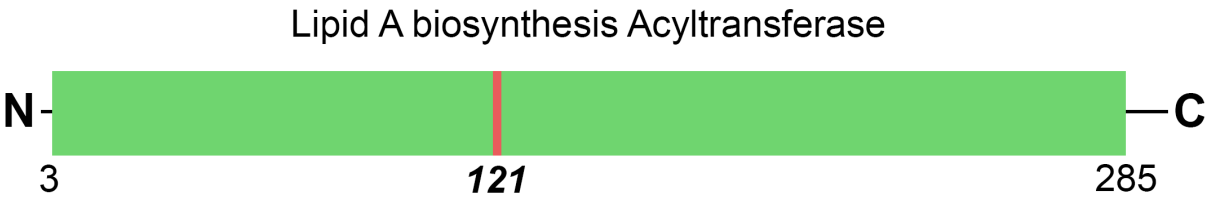

Figure 5.8 – Topology of the CAMP1055/Cj1134 locus, encoding HtrB - an acyltransferase involved in lipid A biosynthesis

Residue numbers provided below domains show regions of Pfam alignment. Residue 121 contained a minor allele variant at the third codon position (Table 5.2).

5.4.1.2.3. Flagellar Loci

Two isolates were found to contain minor allele variation within loci associated with flagella. These isolates: D1-29 and D1-35, belong to CC's ST-48 and ST-464 respectively.

CAMP0303 (Cj0337c) encodes MotA – a putative flagellar motor proton channel protein. Isolate D1-29 - belonging to CC ST-48 – contained a single minor allele variant at residue number 190. This variant is located within a MotA proton channel protein identified by Pfam protein sequence alignment (Figure 5.9). The arrangement of the MotA protein in the greater flagellar motor stator ring complex is displayed in Figure 5.10

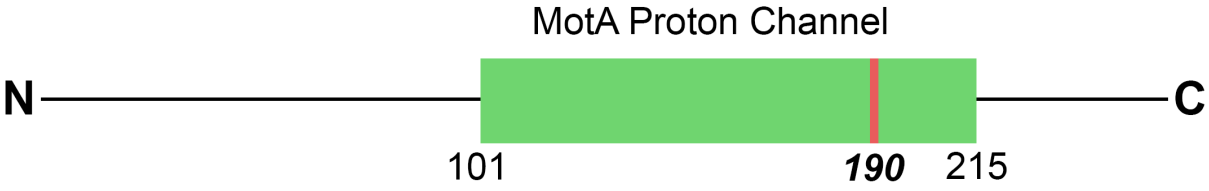

Figure 5.9 - Topology of the CAMP0303/Cj0337c locus encoding the flagellar motor proton channel protein, MotA

Residue numbers provided below domains show regions of Pfam alignment. Residue 190 yielded variation at the second codon position (Table 5.2).

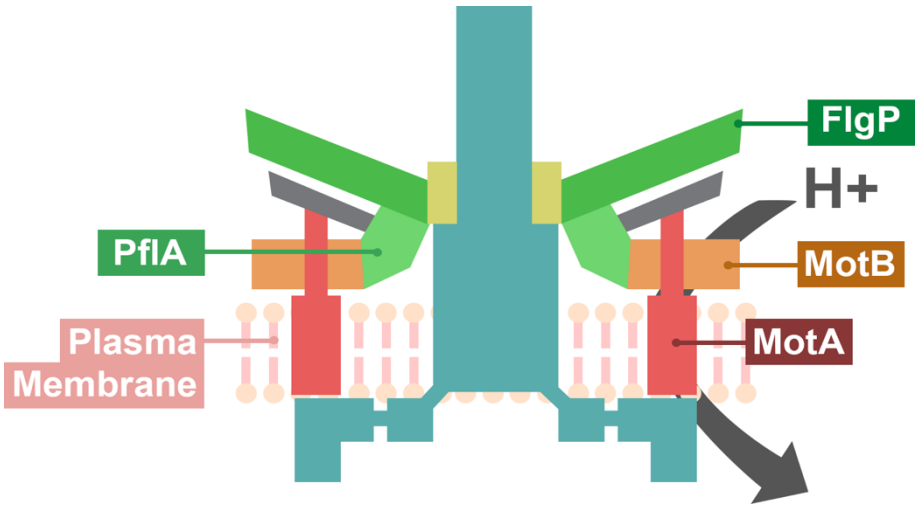

Figure 5.10 –Topology and arrangement of *C. jejuni*'s flagellar motor apparatus

Image adapted from figure 6 in Beeby *et al*., (2016).

CAMP1213 (Cj1295) encodes an aminopeptidase that is linked to flagellar glycosylation. Two minor allele variants were detected at residues 279 and 281, which were mapped to a domain of unknown function with putative peptidase activity (Figure 5.11). The domain of unknown function (DUF4910) showed homology to a putative peptidase identified in *Clostridium acetobutylicum*.

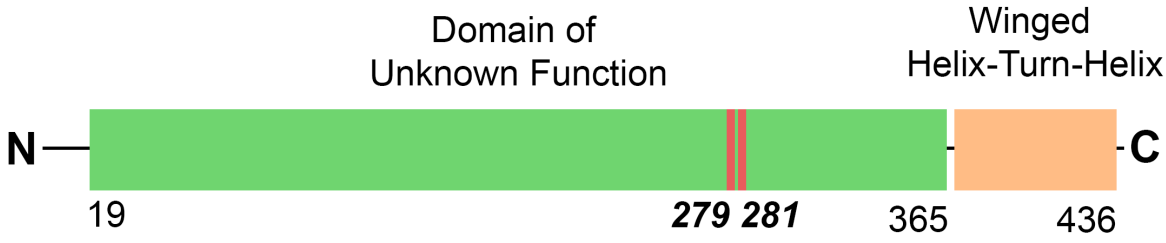

Figure 5.11 - Topology of the CAMP1213/Cj1295 locus

Residue numbers provided below domains show regions of Pfam alignment. Residue 279 and 281 yielded variation at the third and second base position respectively (Table 5.2).

5.4.1.3. Non ST-21 loci relative to NCTC11168

In order to provide a common reference point, loci exhibiting minor allele variation were mapped against the reference assembly of *C. jejuni* strain NCTC11168. The positions of these loci relative to NCTC11168 are provided in Figure 5.12.

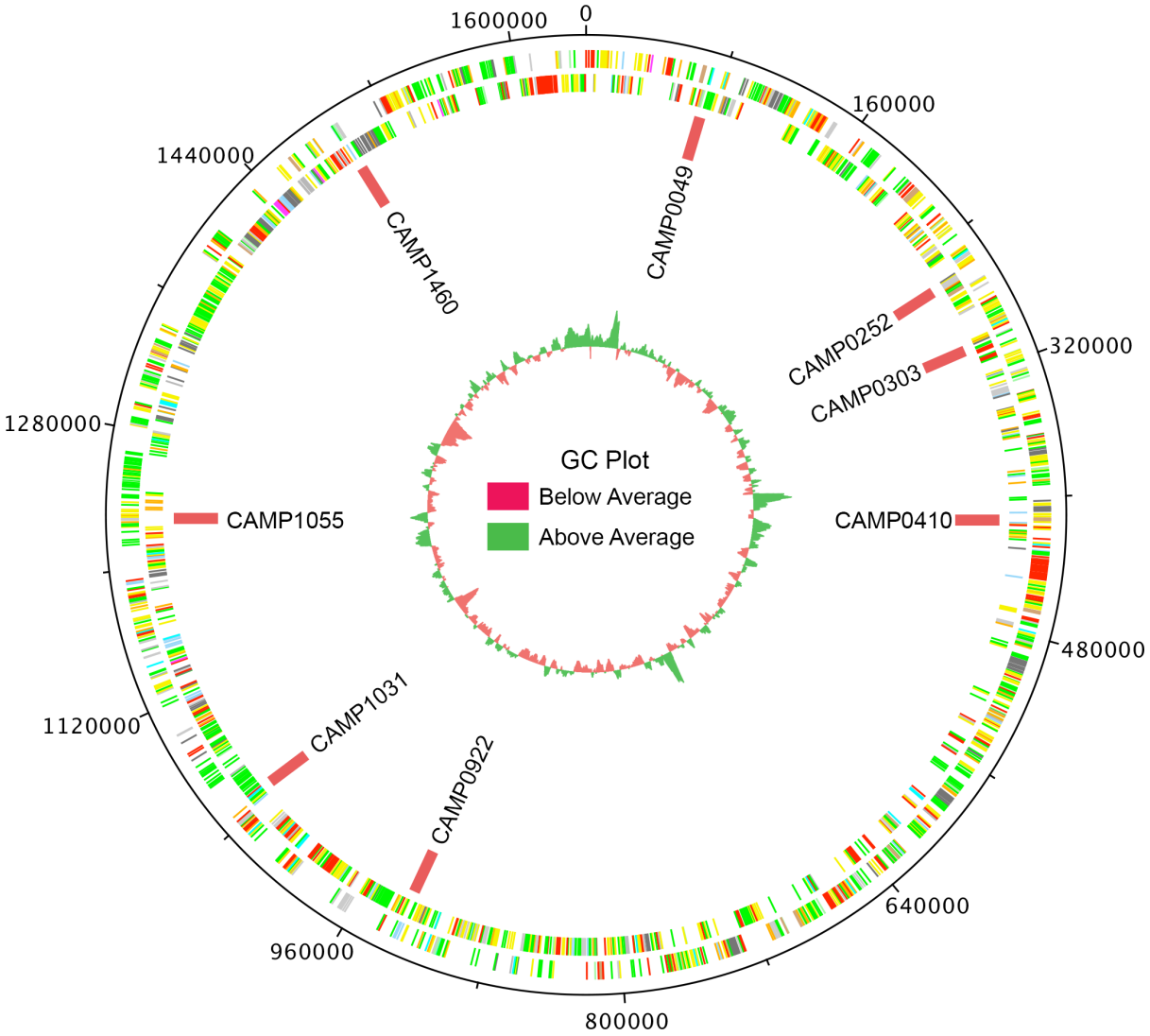

Figure 5.12 - Circular chromosome of NCTC11168 with annotated CDS* and chromosomal positions of variable loci detected within the non ST-21 dataset

CDS as defined in the updated re-annotation of the complete NCTC11168 genome sequence (Gundogdu *et al.* 2007). Highlighted CDS were found to contain minor allele variation at one or more amino acid residues.

5.4.2. Poultry Isolates

5.4.2.1. Isolates, sequencing, and assembly

Out of 12 sequenced isolates, a total of 5 poultry sequences with sufficient quality, purity and depth were produced to an average depth of 164X. These isolates belonged to 3 CC's: ST-257, ST-464 and ST-828 (Table 5.3).

Table 5.3 - Sequencing coverage depth and clonal complex of the five isolates belonging to the poultry deep sequencing dataset

Isolate	Coverage Depth	Clonal Complex
DC1	281	257
DC2	120	257
DC3	281	464
DC5	177	828
DC7	225	828

The poultry deep sequencing dataset had an average N50 of 298,027bp (95% C.I: 197,953bp to 398,100bp), a mean number of contigs of 30 (95% C.I: 7.98 to 52.02), and a mean total length of 1,711,249bp (95% C.I: 1,612,504bp to 1,809,994bp). Figure 5.13 shows the distribution of these quality metrics.

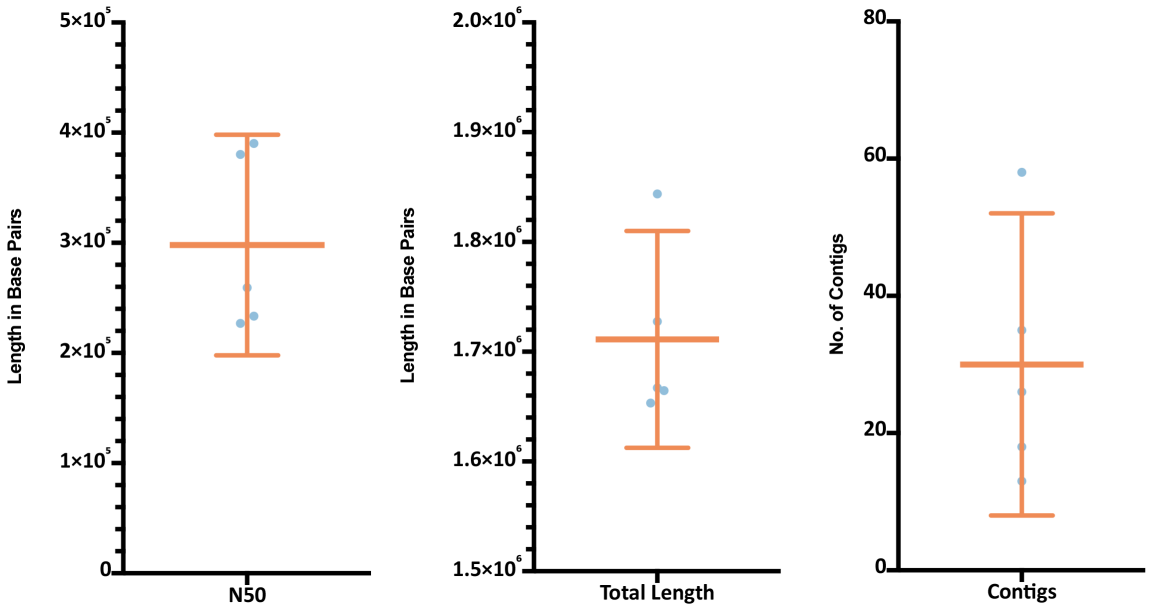

Figure 5.13 - Quality metrics of draft de novo assemblies, with mean inclusive of upper and lower confidence intervals highlighted in orange

Linear regression was used to assess whether the sequencing coverage depth, assembly N50 or number of contigs had an impact on the number of loci in which minor allele variation was detected (Figure 5.14). These analyses revealed no statistical significance (P = 0.2784, 0.1993 and 0.6737 respectively).

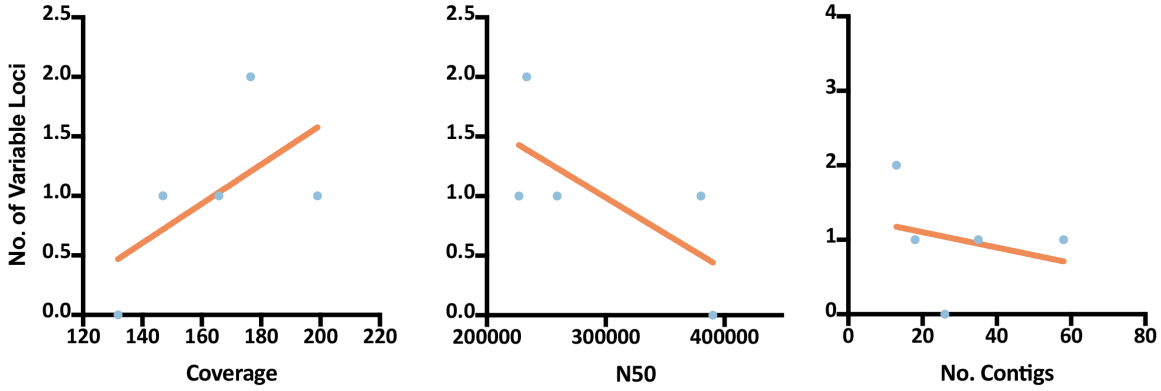

Figure 5.14 - The statistical* impact of sequencing and assembly metrics on the number of loci in which minor allele variation was detected

*Statistical significance calculated using a linear regression analysis, in all instances results were statistically insignificant (Coverage: P = 0.2784, N50: P = 0.1993, No. Contigs: P = 0.6737).

5.4.2.2. Minor allele variation

The poultry deep sequencing dataset contained 10 minor allele variants with sufficient depth and quality to qualify further analysis. Of the 5 isolates sequenced, 2 did not yield any minor allele variants that passed the quality filtering process.

All variants were detected within identified loci. The loci in which minor allele variation was detected, as well as their relative position and effect, were grouped according to the gene product function (Table 5.4).

Table 5.4 - Genes in which minor allele variation was detected within poultry isolates, grouped according to putative function

Isolate	Locus	Functional Group	Residue	Codon			Amino Acid			Mapping Quality*	Coverage Depth	MAF†
				Ref		Alt	Ref		Alt			
DC4	CAMP1196	Membrane	26	TT**T**	→	TT**C**	Phenylalanine	→	Phenylalanine	167	36	0.40
DC1	CAMP1255	Flagellar	263	GT**C**	→	GT**T**	Valine	→	Valine	221	133	0.49
DC1	CAMP1255	Flagellar	271	AA**C**	→	AA**T**	Asparagine	→	Asparagine	221	136	0.48
DC7	CAMP1910	Hypothetical	100	**C**AC	→	**A**AC	Histidine	→	Asparagine	106	193	0.31
DC7	CAMP1910	Hypothetical	100	CA**C**	→	CA**T**	Histidine	→	Histidine	98	216	0.30
DC7	CAMP1910	Hypothetical	164	**G**AG	→	**A**AG	Glutamic acid	→	Lysine	186	378	0.26
DC7	CAMP1910	Hypothetical	164	GA**G**	→	GA**A**	Glutamic acid	→	Glutamic acid	174	376	0.25
DC7	CAMP1910	Hypothetical	171	AG**C**	→	AG**T**	Serine	→	Serine	224	344	0.27
DC7	CAMP1910	Hypothetical	179	GA**A**	→	GA**G**	Glutamic acid	→	Glutamic acid	224	344	0.17
DC7	CAMP1910	Hypothetical	180	**G**CT	→	**C**CT	Alanine	→	Proline	78	260	0.17

* - PHRED-quality scores

† - Minor allele frequency, *i.e.* proportion of reads supporting the alternative allele.

5.4.2.2.1. Membrane

A single isolate contained a minor allele variant within the putative membrane locus CAMP1196. CAMP1196 – also referred to as Cj1277c – encodes a putative ABC transporter protein. This locus has not been characterised, and as such its exact function is unknown. Generally, ABC transporters are a family of membrane bound proteins, and in bacteria are involved in the active transport of molecules across the membrane surface (Garmory & Titball 2004).

Protein sequence alignment to the Pfam database revealed a single conserved ATP binding domain (Figure 5.15), in which a synonymous mutation was discovered at residue number 26.

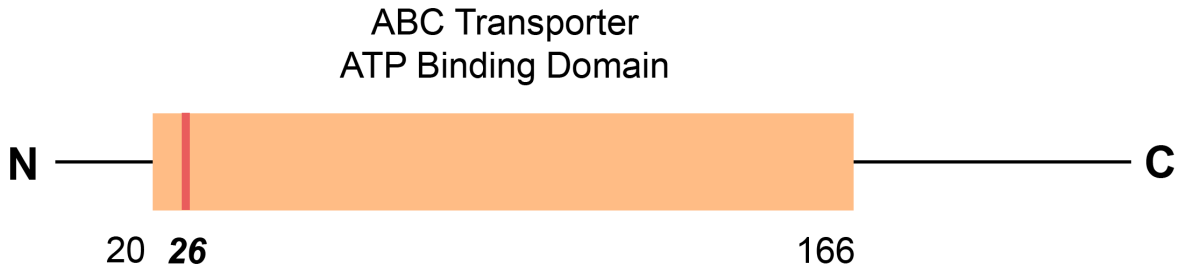

Figure 5.15 - Topology of the CAMP1196/Cj1277c locus

Residue numbers provided below domains show regions of Pfam alignment. Residue 26 yielded variation at its third base position, shown in Table 5.4.

5.4.2.2.2. Flagellar

A single isolate was found to contain two minor allele variations within the flagellar locus CAMP1255. This locus is commonly referred to as *flaA*, and encodes one of two flagellin subunits present in *Campylobacter spp*. (Power *et al*. 1994). FlaA is the predominant flagellin protein, and has been shown to be essential for cellular invasion (Wassenaar *et al*. 1991).

This locus was found to contain two synonymous variations occurring at intermediate frequency within a flagellar hook motif (Figure 5.16).

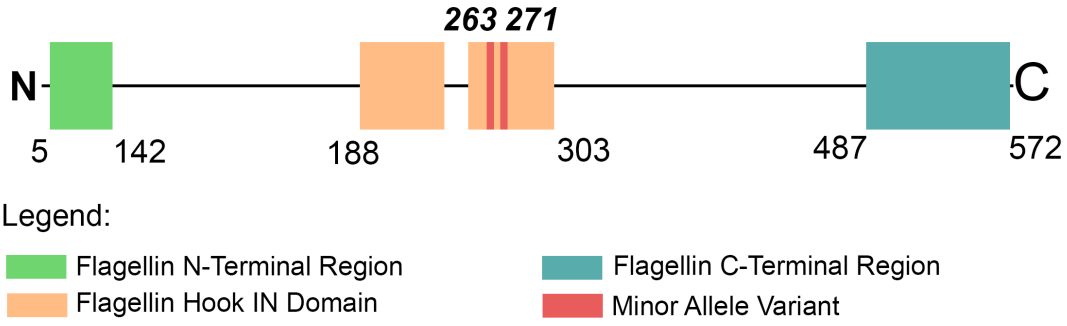

Figure 5.16 - Topology of the *flaA* flagellin locus

Residue numbers provided below domains show Pfam identified alignments.
Residue 263 and 271 yielded a minor allele variant at the third base position (Table 5.4).

5.4.2.2.3. Hypothetical

Isolate DC7 contained 7 minor allele variants within the CAMP1910 locus. This locus encodes a hypothetical protein exclusively found in *C. coli*, and contains no conserved domains or assigned putative function.

The number of variants detected within this locus may be spuriously represented. Residue 100 was found to have two polymorphic bases in the first and third codon positions. Given their similar minor allele frequencies, these variants may in fact be representative of a single alternative codon, which in this instance would be asparagine (AAT).

5.4.2.3. Poultry loci relative to NCTC11168

In order to provide a common reference point, loci exhibiting minor allele variation were mapped against the complete published assembly of *C. jejuni* strain NCTC11168 (J Parkhill *et al.* 2000).

Where possible, the positions of these loci relative to NCTC11168 is provided in Figure 5.17. CAMP1910 is a putative protein only identified in *C. coli* and as such could not be mapped to the 11168 chromosome.

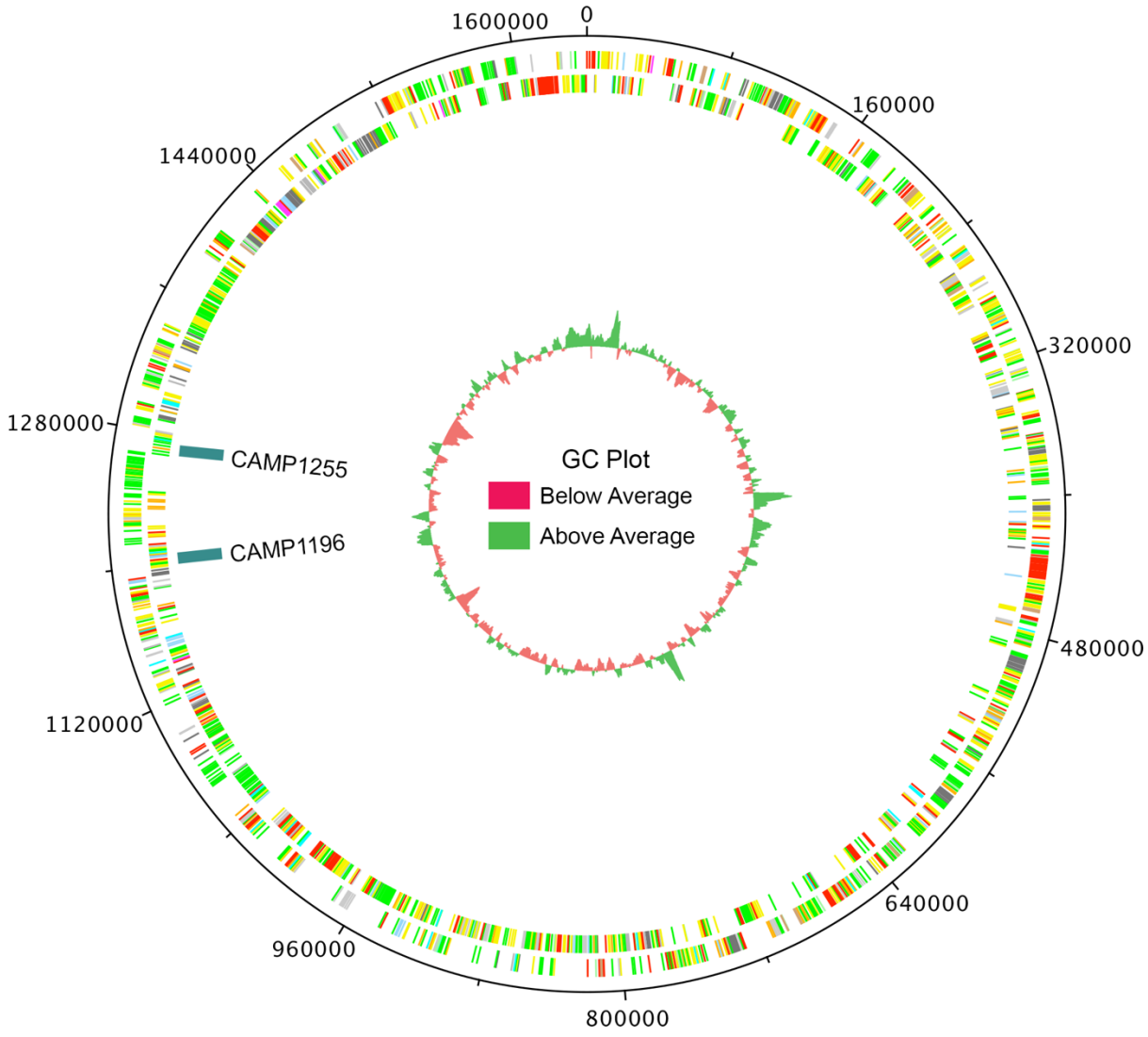

Figure 5.17 - Circular chromosome of NCTC11168 with annotated CDS* and chromosomal positions of variable loci detected within the Poultry dataset

CDS as annotated in the updated re-annotation of the complete NCTC11168 genome sequence (Gundogdu *et al.* 2007). *Highlighted CDS were found to contain minor allele variation at one or more amino acid residues.

5.4.3. Cumulative Statistics

Due to the different sizes of the three datasets (*i.e.* ST-21, Poultry and non ST-21), and the large variance between assembly metric distributions (Figure 5.18), the datasets were tested as a whole to ensure there was no statistically significant difference between them. The coverage, N50 and total number of contigs were considered along with the number of variable loci detected. Using linear regression, no statistical difference was detected (Figure 5.19).

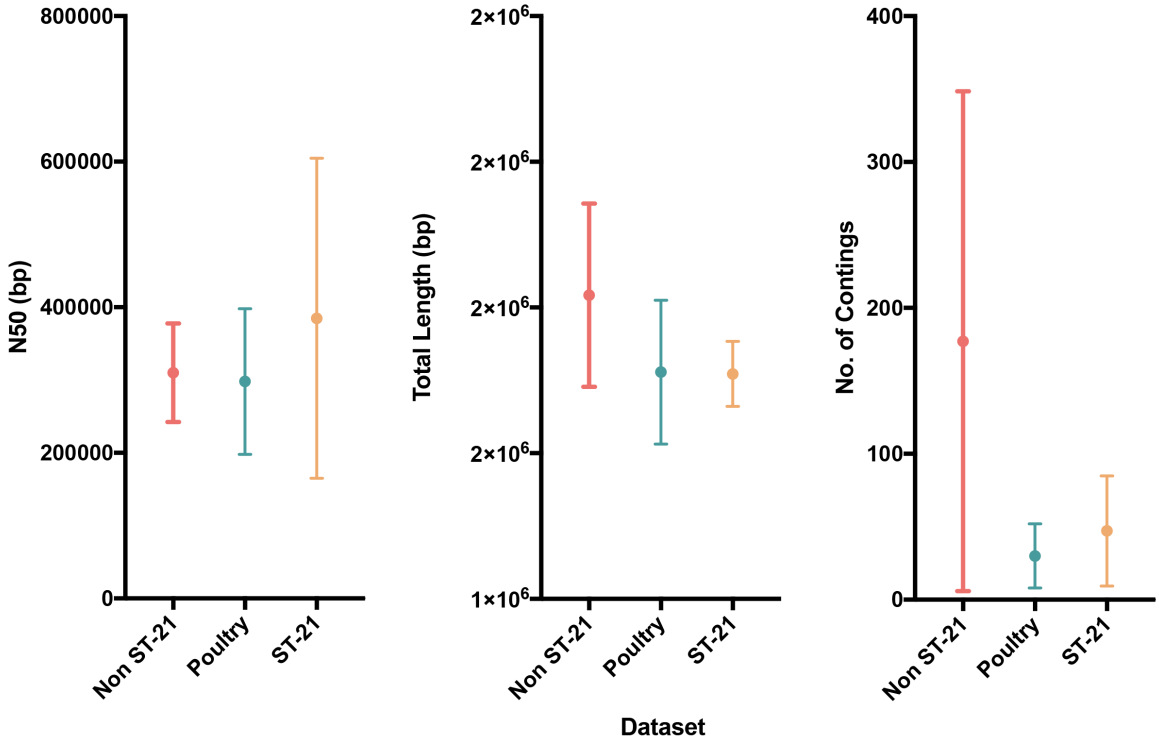

Figure 5.18 - Distribution of assembly metrics between the three deep sequencing datasets

Non ST-21 (N=11), Poultry (N=5) and ST-21 (N=17).
Point provides the mean, with whiskers displaying the 95% confidence interval.

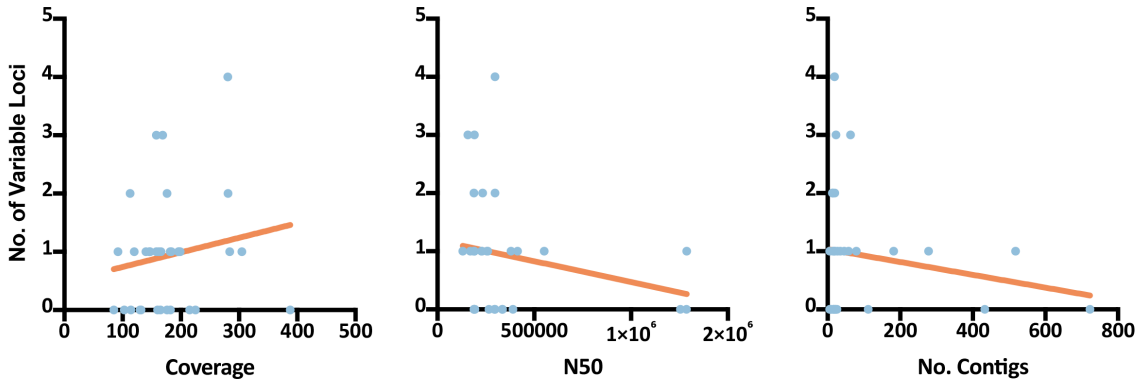

Figure 5.19 - The statistical* impact of sequencing and assembly metrics on the number of loci in which minor allele variation was detected

*-Statistical significance calculated using a linear regression analysis, in all instances results were statistically insignificant (Coverage: P = 0.3558, N50: P = 0.2114, No. Contigs: P = 0.3105)

5.4.4. Comparison of Diversity

To allow comparison between datasets, key indicators were calculated based on the average number of loci, variants, intergenic variants, and percentage of each data set that did not yield any variation Table 5.5. Due to the differences in the number of isolates constituting each data set, these values are provided as average number per isolate.

Table 5.5 - Measures of the diversity observed in three deep sequencing datasets, normalised to account for the variance in sample size

	Dataset		
	Non ST-21	ST-21	Poultry
Variable Loci*	0.82	0.35	0.60
Minor allele variants†	1.80	2.11	1.60
Intergenic variants	0.18	0.24	0.00
Grouped locus functions‡			
Chemotaxis variants	1.00	0.71	0.00
Membrane variants	0.36	0.35	0.20
Flagellar variants	0.27	0.53	0.20
Total	1.63	1.59	0.40
Syn/Non-syn ratio§			
Chemotaxis variants	0.45	0.15	0.00
Membrane variants	0.75	0.17	1.00
Flagellar variants	0.33	0.30	1.00
Total	0.50	0.21	1.00
No variation detected‖	36%	35%	40%

* - Number of loci identified as containing variants on a per isolate basis (*i.e.* loci containing variants in multiple isolates amongst a dataset are counted multiple times).
† - All variants, inclusive of intragenic and intergenic positions.
‡ - Variants are grouped according to putative gene product function.
§ - The ratio of variants observed to cause synonymous vs. non-synonymous changes to the amino acid residue, again grouped according to putative gene product function.
‖ - Percentage of the isolates amongst datasets that did not yield any minor allele variants.

5.4.5. All loci relative to NCTC11168

In order to ease comparison between the Non ST-21 and poultry datasets presented within this chapter and Chapter 4's ST-21 dataset, variable loci were mapped to the NCTC11168 reference genome (Figure 5.20) where possible. Two loci were found to contain minor allele variants in both the Non ST-21 and ST-21 datasets. Intergenic variants (including those contained within intergenic homopolymeric tracts) as well as the *C. coli* specific locus CAMP1910 are not shown.

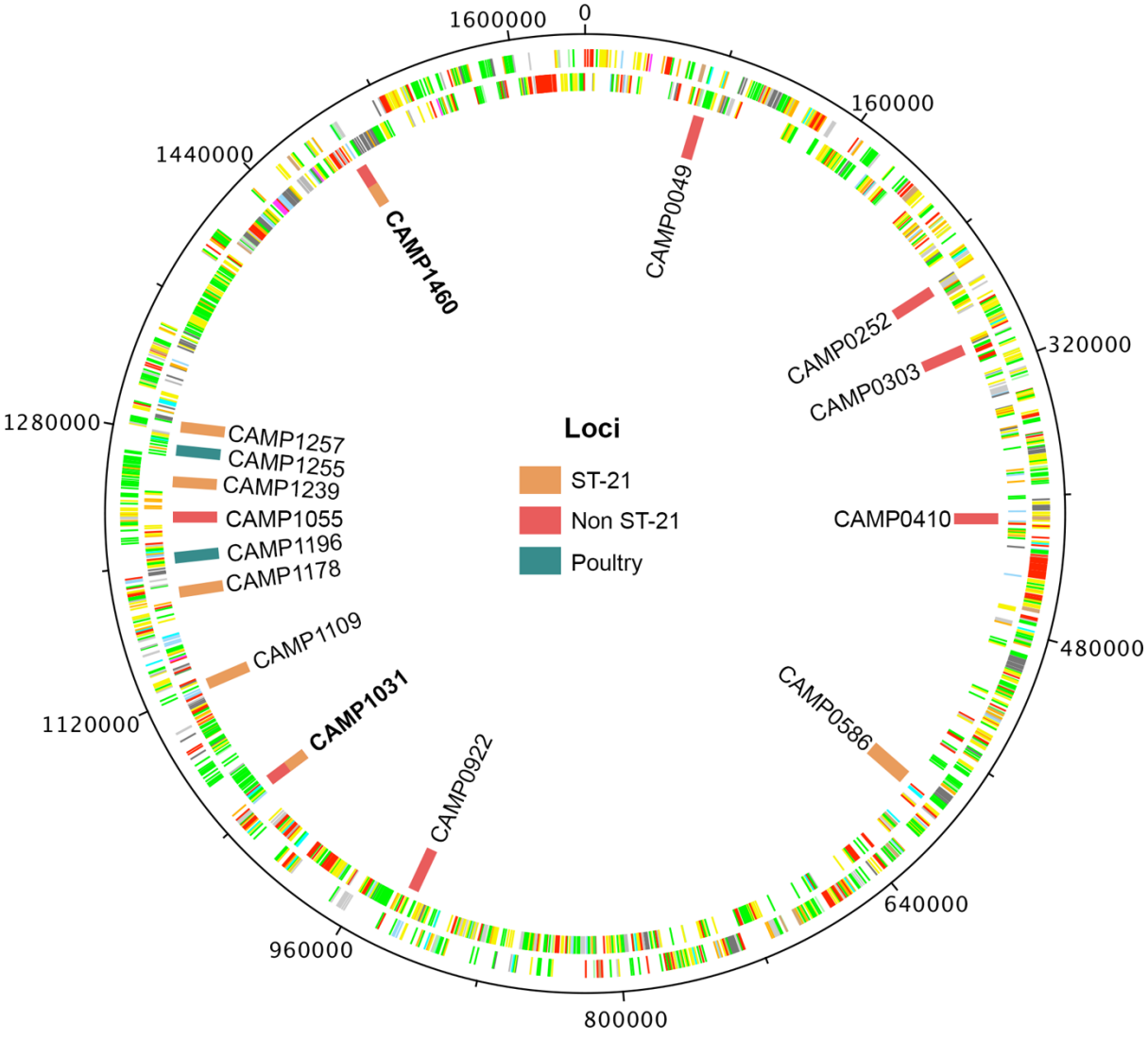

Figure 5.20 - Circular chromosome of NCTC111168 with approximate* locations of variable loci detected across all deep sequencing isolates

*Some loci are shifted from their precise location to increase legibility.
CDS as annotated in the updated re-annotation of the complete NCTC111168 genome sequence (Gundogdu *et al.* 2007). Highlighted CDS were found to contain minor allele variation at one or more amino acid residues.
Loci highlighted in bold indicate a locus that was found to be variable in multiple datasets.

Table 5.6 - Locus tag*, function and known aliases of loci containing minor allele variation across all deep sequencing datasets

Locus Tag*	Functional Group	Locus Aliases	Datasets†		
			Non ST-21	ST-21	Poultry
CAMP0049	Membrane	*cj0057*			
CAMP0252	Chemotaxis	*cj0283c*, *cheW*			
CAMP0303	Flagellar	*cj0337c*, *motA*			
CAMP0410	Chemotaxis	*cj0448c*, *Tlp6*			
CAMP0586	Membrane	*cj0628*, *capA*			
CAMP0922	Membrane	*cj0999c*			
CAMP1031	Chemotaxis	*cj1110c*, *cetZ*			
CAMP1055	Membrane	*cj1134*, *htrB*			
CAMP1109	Chemotaxis	*cj1190c*, *cetA*			
CAMP1178	Membrane	*cj1259*, *porA*			
CAMP1196	Membrane	*cj1277c*			
CAMP1213	Flagellar	*cj1295*			
CAMP1255	Flagellar	*cj1339c*, *flaA*			
CAMP1257	Flagellar	*cj1341c*, *maf6*			
CAMP1460	Chemotaxis	*cj1564*, *CcmL*			

* - As defined in the PubMLST *C. jejuni* and *C. coli* database (Jolley & Maiden 2010)
† - Datasets found to contain variation within specified locus, as shown in Figure 5.20.

The common reference tag of all loci identified as variable within the deep sequence data, as well as their common aliases and putative function are provided in Table 5.6.

Figure 5.21 shows synonymous and non-synonymous variants detected across all of the deep sequence data within loci mapped to the NCTC11168 genome. This reveals a higher number of non-synonymous mutations in the ST-21 dataset across all three functional groupings of loci. The non ST-21 dataset yields a more diffuse pattern of diversity, with variants spread across several loci.

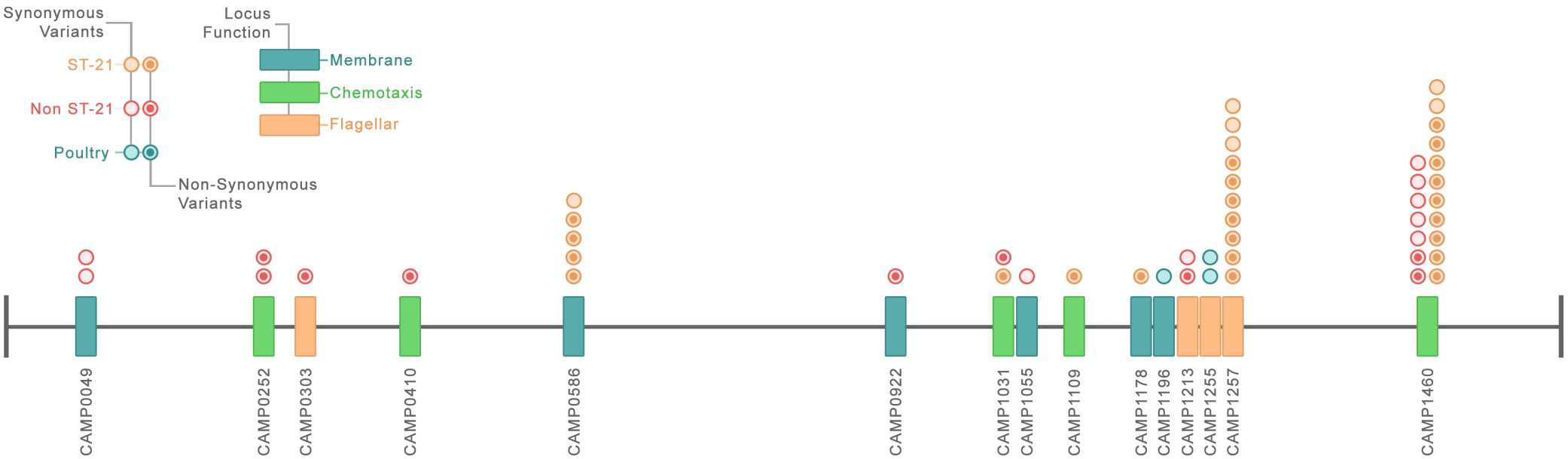

Figure 5.21 – Loci containing synonymous and non-synonymous minor allele variants across all deep sequencing datasets.

Line representing a linearised *Campylobacter* chromosome with genes displayed in their relative positions and coloured according to product function. Variants detected within the dataset are displayed as circles above the genes, with non-synonymous variants containing a filled centre and synonymous variants hollow. Variants are also colour coded to represent their constituent clonal complex.

5.5. Discussion

The findings presented in this chapter show that intra-population diversity is not unique amongst the ST-21 clonal lineage. In fact, many *C. jejuni* complexes, and the *C. coli* complex ST-828 also contain variation within a single infectious population. Additionally, minor allele variation is present amongst populations isolated from fresh retail poultry, albeit in a reduced capacity.

To aid in the navigation of this discussion, it is important to note the three datasets that are considered. Namely: the clinical 'non ST-21' and the fresh 'poultry' datasets presented within this chapter, and - as the results are inherently linked - the results from the clinical 'ST-21' dataset presented in Chapter 4.

5.5.1. Non ST-21 Isolates

This dataset expands on the work presented in the previous chapter by including a variety of additional clonal lineages. The sampled isolates represent a further five CC's from clinical infections: ST-828, ST-48, ST-353, ST-464 and ST-45. These complexes were among the most abundant lineages identified in Chapter 3's genetic epidemiology dataset, and represent examples of lineages that are successful human pathogens.

5.5.1.1. Chemotaxis Genes

In both this dataset, and Chapter 4's ST-21 dataset, chemotaxis genes are the most common functional grouping of loci that are found to undergo minor allele variation. The one commonality that both sets of isolates share is that they are of clinical origin, and as such represent populations that were successful in propagating a human infection. Furthermore, these isolates originate from cases that occur with a severity that is sufficient to warrant medical intervention. In the United Kingdom, it is estimated that 1 in 9.3 cases of campylobacteriosis go unreported (Tam *et al.* 2012), and as such the constituent clinical cases represent isolates that

generate a greater level of pathogenicity compared to the wider *Campylobacter spp*. population as a whole. The repeated variation within genes pertaining to chemotaxis functions would suggest that these loci are coming under strong selective pressure within the context of the human gut environment.

The intragenic location of the detected minor allele variants was resolved using amino acid sequence alignment to the Pfam protein database (Finn *et al*. 2016). This revealed that in several of the identified chemotaxis loci (*i.e*. CAMP0410, CAMP1031, CAMP1460) minor allele variation was located within a methyl-accepting (MA) signalling domain. This domain is conserved across all of *Campylobacter*'s ligand sensing transducer like proteins (Tlps) and is responsible for signal transduction as part of the greater chemotaxis response (discussed in greater detail in Chapter 1). Briefly, the MA-signalling domain acts as a binding site for the scaffolding protein CheW, which forms a quaternary structure with histidine kinase CheA. This ultimately leads to phosphorylation of CheY, which mediates directional switching of the flagellar motor apparatus (Korolik *et al*. 2008, Zautner *et al*. 2012).

Variation was also discovered within the CheW scaffolding protein itself (CAMP0252). The diversity present in the *cheW* locus may represent a different mechanism that is achieving the same end result as variation in the MA-signalling domain of transducer like proteins (*i.e*. altered sensitivity to flagellar motor switching in response to specific ligands). By altering the binding affinity of the MA-signalling domains (*e.g*. causing a conformational change to the CheW receptor site), or indeed by altering the scaffolding proteins themselves, the response to a given Tlp and its ligands may be modulated. This could for example occur in response to the colonisation of a novel host, where the abundancy and distribution of ligands may be different to that of the originating host species.

One isolate (D1-29) expressed variation at both the *cheW* locus, and in the MA-signalling domain of Tlp8 (CAMP1031) and Tlp6 (CAMP0410), which

reinforces the proposition that chemotaxis loci – and specifically the signal transduction pathway - are coming under strong selective pressures within the human host. This particular isolate belongs to the ST-48 lineage, which is strongly associated with beef (Rotariu *et al.* 2009, Kärenlampi *et al.* 2007). ST-48 is also the 5th most common CC identified in genetic epidemiology studies of UK *Campylobacter* populations (Cody *et al.*, 2012), and the fifth most abundant lineage amongst isolates submitted to the PubMLST *C. jejuni* and *C. coli* database (4.94% of total). The lineage's abundance in beef, and relative success as a human pathogen may in part be due to its ability to adapt to multiple host environments, with the identified variation potentially acting as a driving mechanism by which novel host colonisation occurs.

In addition to the commonality of the functional group as a whole, two of the chemotaxis loci – CAMP1031 and CAMP1460 - were also found to contain variation within the previous chapter's dataset. These loci encode transducer like proteins (Tlps) belonging to groups C and A respectively. The former group represents genes that are relatively uncharacterised, and as such, in depth discussion is limited. In contrast, the CAMP1460 locus has undergone extensive study. The CAMP1460 locus encodes Tlp3 , which has been shown to interact with several chemo-attractants simultaneously, providing its alternative alias, CcmL, or *Campylobacter* chemoreceptor for multiple ligands (Marchant *et al.* 2002, Rahman *et al.* 2014).

Analysis of the BigsDB database for *C. jejuni* and *C. coli* isolates reveals 683 unique CAMP1460 alleles. Of these alleles, only 247 have associated sequence data. The remaining alleles with no sequence data may be a result of automatic processing of submissions before manual curation and subsequent exclusion. The 247 alleles are distributed across 361 isolate sequences. The majority of these alleles are unique, with 76.5% (n=189) of alleles only observed in a single isolate. In comparison, 45% of alleles constituting the housekeeping locus *pgm* are unique, with a much larger sample size (10,688 isolate sequences). This would suggest that the CAMP1460 locus is inherently highly variable. When considering the

main sources of isolates containing a confirmed allele for the CAMP1460 locus, 60.94% are of human origin (N=220), which further supports the hypothesis that this locus – and other similar chemotaxis loci – undergo selection in the human gut.

The CAMP1460 locus has also been shown to be variable in other genomic studies. Results from passaging *C. jejuni* isolate 11168 in mice yielded variation in the CAMP1460 locus at codon 59 (Jerome *et al.* 2011). This was not investigated, as the frequency occurred at <5% (4.9%), however it does support the observation that this locus undergoes variation during the host infection process. The variable codon was not in the same domain as variants identified in the ST-21 and non ST-21 datasets, and instead is located within the extracellular ligand binding domain. This could reflect a difference in host adaptation to mice rather than humans. This poses the question as to whether the variation observed within the MA-signalling domain of CAMP1460 (and other Tlps) is not only a product of human infection and its relative selective pressures, but perhaps specific to the human host entirely.

Evidence for the validity of this is provided in the analysis of an accidental infection of a laboratory worker, with the same *C. jejuni* reference isolate 11168 (Thomas *et al.* 2014). This study again revealed variation within the CAMP1460 locus following human passage, with a single polymorphic base conferring an amino acid substitution at residue 511. In this instance, the variable residue is located within the methyl-accepting signalling domain and exists in close proximity to the minor allele variants detected in the clinical deep sequencing datasets. The researchers in this study went on to passage the strain isolated from the laboratory worker in interleukin-10 deficient mice, which resulted in a reversion of the observed polymorphism. The researchers conclude that the variation at residue 511 may be specific to the human host, which has important implications for the highly similar variations detected within the clinical datasets. It would be interesting to passage the clinical isolates exhibiting variation within chemotaxis loci in mice to see if these

variants also revert, and whether the variation observed is indeed specific to the human host as these results would suggest.

5.5.1.2. Membrane and Flagella

Several variants were detected within loci encoding gene products related to *Campylobacter*'s membrane and flagellar. Notably, variation was found in the locus encoding MotA (CAMP0303), which assembles as part of the larger flagellar stator ring apparatus that ultimately mediates flagellar rotation (Beeby *et al.* 2016). The variable residue that undergoes a non-synonymous mutation is located within one of three predicted transmembrane regions (Mohawk *et al.* 2014), which would be likely to affect protein function, and specifically may modulate proton transport across the membrane. This would result in altered motor function, and may reflect an additional mechanism by which a modulated chemotactic response to ligands is achieved as discussed in relation to the chemotaxis loci.

The *motA* locus has been identified as undergoing significant variation in other genetic studies. Resequencing of a well characterised outbreak strain, 81-176, revealed multiple instances of genetic variation amongst colony variants (Mohawk *et al.* 2014). These colonies displayed alternate phenotypes that varied in opacity and mobility. In future studies, re-culture of this isolate to screen for variant colonies may reveal a phenotype for the genetic variation observed, which would aid in determining the biological effect of the minor allele variant.

Two of the three variable membrane related loci (CAMP0922 and CAMP0049) are uncharacterised, and as such little information is available on the protein function or structure. However, the bacterial membrane is a notable source of genomic variation within *Campylobacter spp.* in general (Cody, Maiden, and Dingle 2009; Q. Zhang *et al.* 2000). The final membrane locus (CAMP1055) encodes the putative acyltransferase, HtrB, which is involved in lipid A biosynthesis. Lipid A is one of two constituent parts of *Camplyobacter*'s lipo-oligosaccharide. This locus has been linked to cell wall formation, and is

essential for cellular survival in stress conditions (Phongsisay *et al.* 2007). *C. jejuni* LOS structures have been shown to be highly diverse, arising from differences in the moieties and linkages of individual sugars that are conferred via transferases within the LOS operon (Gilbert M *et al.* 2008). The *htr*B locus has been proposed to add two secondary acyl groups to the lipid A molecule, and as such variation within this locus may result in an altered LOS phenotype.

Loci pertaining to membrane and flagellar functions were also identified with a similar relative frequency in the ST-21 clinical dataset, which - as is seen in the case of chemotaxis loci – suggests these functional groups are undergoing selection within the human gut. Several membrane and flagellar gene products are shown to be highly immunogenic (Baqar *et al.* 2008, Poly *et al.* 2007a, Shoaf-Sweeney *et al.* 2008, Burnens *et al.* 1995, Nielsen *et al.* 2012), which may explain the variation observed.

5.5.1.3. *C. coli*

Four isolates analysed within the non ST-21 dataset belong to the *Campylobacter coli* CC, ST-828. When comparing the variation within ST-828 isolates to other non ST-21 *C. jejuni* isolates, a similar pattern of diversity can be observed. Again, the most common functional grouping of loci (in order of abundance) encode chemotaxis, membrane and flagellar associated proteins.

The ST-828 isolates yielded variation in three loci: CAMP1460, CAMP0252, and CAMP0049. CAMP1460 and CAMP0252 are involved in chemotaxis, and were also found to contain variation in clinical *C. jejuni* isolates. Notably, CAMP1460 represents a locus that has been identified as containing variation across multiple *C. jejuni* CC's, and now also across multiple species of *Campylobacter* (*i.e. C. coli*). This further emphasises the frequency at which this locus is encountering selective pressure within human hosts.

The final locus (CAMP0049) is exclusively found in *C. coli* isolates. Whilst the gene itself is uncharacterised, the presence of diversity at this locus

highlights the importance of including ST-828 isolates in the non ST-21 dataset. Further study of the genetic heterogeneity of *C. coli* may reveal additional species-specific loci that are involved in the generation of clinical disease that are yet to be elucidated.

5.5.2. Poultry Dataset

The purpose of this dataset was to provide an insight into the level of genetic heterogeneity amongst pre-infectious *Campylobacter* populations. As poultry is a major source of human campylobacteriosis (Wilson *et al.* 2008), these isolates represent typical isolates that the public may encounter, and which ultimately may generate an infection. The poultry isolates all belong to CC's that are abundant in human disease, and as such are an appropriate sample to compare with the clinical datasets.

Isolating *Campylobacter spp*. from retail poultry meat proved to be challenging, due to significant contaminant growth. Several variations on the culturing methodology were tested for their efficacy at reducing contamination, including the addition of a supplement intended to increase the aerotolerance of *Campylobacter spp.,* as well as altering the temperature and duration of incubations. However, these variations had little discernible impact and ultimately, it was not possible to consistently produce cultures of pure *Campylobacter spp*. colonies. In the majority of instances, samples were obtained from one of many duplicate plates with the least contaminant growth, and with sufficient separation between morphologically distinct colonies. When transferring bacterial growth for DNA isolation, great effort was taken to ensure that the loop did not make contact with contaminants, whilst still sampling all present *Campylobacter* colonies. Despite this, contaminant reads were detected in the final assembly of three additional poultry isolates.

To test whether reads from contaminant bacteria would impact the calling of minor allele variation, a mock read community was constructed *in silico*. Briefly, this involved the random interspersion of *Campylobacter spp*. genomic reads with those from example

contaminant organisms *E. coli* and *S. enterica* in various relative abundancies. The main issue revealed with contaminant reads is in the generation of a *de novo* assembly, with increasingly poorer assemblies constructed as a result of increasingly contaminated read files. As the deep sequencing methodology relies on the alignment of an isolate's genomic read data to their resulting *de novo* assembly, contaminant reads created a significant barrier to analysis.

With this in mind, contaminated poultry read files were excluded from analysis, and as such the poultry dataset is smaller than anticipated. However, it does provide preliminary evidence that shows clinical populations of *Campylobacter spp.* exhibit a different pattern of intra-population diversity to poultry isolates.

5.5.2.1. Genetic Diversity

The poultry dataset exhibited less overall diversity than that observed in clinical isolates, as reflected by the number of variable loci, minor allele variants and functional variants per isolate. This is reasonable when considering the relatively harsh environment retail poultry meat offers. Studying the survival of *Campylobacter jejuni* on refrigerated chicken meat revealed it is able to endure long periods of time at 4°C, though populations exhibit a reduced cell count (Bhaduri & Cottrell 2004). A reduction of the total *Campylobacter* population may create a genetic bottleneck and lead to a smaller number of dominant clones that persist and are recoverable during sampling. A similar bottleneck is likely to be experienced during slaughter and preparation, either as a consequence of the techniques used to handle the carcass (*e.g.* during feather scalding) or via direct interventions used to reduce bacterial contamination (*e.g.* carcass rinses). The combination of these reductive events may explain the decrease in diversity observed amongst poultry isolates.

A notable difference amongst the poultry dataset was the lack of variation identified within chemotaxis loci, which represented the dominant functional grouping of variable alleles in both clinical groups.

This is logical, given the abundance of chemotactic ligands would be relatively stable and homogenous on poultry meat compared to the human gut environment, and as such chemotaxis would confer less of an advantage to the resident *Campylobacter* population. Additionally, *Campylobacter spp*. would not be actively growing and as such would have no requirement for chemotaxis for nutritional purposes. These results further support the conclusion that variation within chemotaxis loci occurs as a result of selective pressures exerted by the human host.

Campylobacter spp. have been shown to produce viable but non-culturable (VNC) cells under stress conditions (Moore 2001, Murphy *et al*. 2006). Many of these stress conditions would be exerted on isolates colonising retail poultry meat, including lack of nutrient availability and low temperature (Murphy *et al*. 2006, Lázaro *et al*. 1999). Passaging this stress-associated form of *Campylobacter spp*. in mice resulted in a symptomatic infection, and saw a reversion to a culturable state (Saha *et al*. 1991). As the DNA analysed in the deep sequencing dataset is extracted from fresh *Campylobacter* plate cultures, any cells that are present in a viable but non-culturable form would not be represented. However, following passage in a human host, these isolates may later recover and revert to a culturable form that can be sampled. This may suggest that variable loci identified exclusively in the clinical datasets are not necessarily invariable in the poultry isolates. Rather, the sampling methodology is limited by its reliance on culturable forms of *Campylobacter spp*.

All variants detected within the poultry dataset were synonymous, which was not the case with either clinical group. Whilst synonymous mutations have less of an immediately apparent effect, emerging research shows that there is still an impact associated with alternate codon usage. During translation, amino acid sequences are read via isoaccepting tRNAs that recognise specific codons. The concentration of these tRNAs has been shown to be highly variable (Dong *et al*. 1996), and as such translation of a synonymous codon with fewer complimentary tRNAs is found to occur at a slower rate, which can ultimately influence

the efficiency of protein folding (Zhang *et al.* 2009). Synonymous mutations have also been shown to affect the level of transcription (Agashe *et al.* 2016). With this in mind, the synonymous mutations observed within the poultry loci may have modulated their rate of transcription, potentially in response to the stressful conditions present on retail poultry meat (*e.g.* temperature, amino acid availability). The bias towards synonymous mutations may also be reflective of alternative codon usage in the chicken gut, which is later undergoing variation as a result of novel host infection.

5.5.3. Comparing deep sequence datasets

5.5.3.1. Distribution of Loci

Figure 5.20 provides the approximate positions of all variable loci detected amongst the deep sequencing isolates. The majority of loci appear to be distributed equally across the 11168 chromosome, however there is one region that is found to contain numerous loci. This region, starting with the CAMP1178 locus at chromosomal position 118921 and ending with the CAMP1257 locus at position 1274684, contains 40% of the variable loci identified across all datasets. All of the genes contained within this region are functionally related to either the flagella or membrane, which is a noted area of hyper-variability within the 11168 genome (J Parkhill *et al.* 2000), and contains gene products that have shown to illicit a highly immunogenic response (Burnens *et al.* 1995, Nuijten *et al.* 1991, Huang *et al.* 2007).

5.5.3.2. Differences in the clinical datasets

The non ST-21 dataset exhibited a decreased number of variants per isolate compared to ST-21 isolates (~1.80 vs ~2.11 variants per isolate respectively, Table 5.5). Conversely, the non ST-21 variants were distributed across more than double the number of ST-21 loci, when factoring in the difference in dataset size (0.82 vs 0.35 respectively). This may suggest that the ST-21 lineage has a gene repertoire with specific alleles that are more adapted to the human host than non ST-21 isolates, and as a result do not experience variation is as many genes in response

to selective pressure. This would make sense given that ST-21 is the most commonly identified CC in human infection.

When considering the functional grouping of loci in relation to the observed variation, both ST-21 and non ST-21 isolates contain approximately the same number of per-isolate variants (1.59 and 1.63 repsectively). This disparity between total variation, and variation specific to the chemotaxis, membrane and flagellar genes is mainly due to the presence of variation within a single homopolymeric tract identified within several ST-21 isolates. This tract mediates the expression of a flagellar glycosylation locus (CAMP1239, discussed in Chapter 4), and was identified as variable in 23% (n=4) of the ST-21 dataset. The non ST-21 dataset did not reveal any variation within this tract - which given its abundance in ST-21 isolates - would suggest the observed variation is exclusive to the ST-21 CC.

There are also important differences observed between the two clinical datasets in relation to the most frequently identified variable loci: CAMP1460. This locus is found to contain variation within 29% (n=5) of ST-21 isolates, compared to 18% (n=2) of non ST-21 isolates. Further, there is a notable difference in the number of synonymous and non-synonymous variants between these groups. ST-21 isolates yielded a total of 11 minor allele variants, of which ~82% (n=9) were non-synonymous. In comparison, non ST-21 isolates yielded 7 variants, with just 29% resulting in a non-synonymous substitution. This may also be compounded by the fact that a single isolate (D7-17) yielded two minor allele variants at positions 1 and 2 of amino acid residue 606. Both variants occurred at a minor allele frequency of 0.48, and therefore likely occur concurrently in a single haplotype. This would result in the reference serine codon AGT undergoing a synonymous mutation to alternative serine codon TCT, rather than the two non-synonymous mutations detailed in Table 5.2.
The bias in ST-21 variants towards non-synonymous mutations is also apparent in other chemotaxis and membrane loci, and indeed in comparison of the functionally grouped loci as a whole, with 79% of all

ST-21 variants resulting in non-synonymous changes vs. 50% in non-ST-21 complexes. These results indicate that ST-21 isolates are encountering a stronger selective pressure during human infection, and undergoing an increased level of genomic variation in a smaller range of loci as a result. This is not to say that the synonymous mutations are not important – as discussed above, synonymous codon usage impacts transcription rate and protein expression (Dong *et al.* 1996, Zhang *et al.* 2009). It is also important to consider the typical duration of a symptomatic *Campylobacter spp.* infection, which lasts for approximately 2-7 days. This relatively short period of exposure is generating large amounts of diversity, with both synonymous and non-synonymous effects.

Across all datasets, the proportion of isolates that did not reveal any minor allele variation is highly similar (~37%). This does not necessarily mean that these isolates did not contain any variation. Rather, the stringent quality filtering employed by the analysis pipeline did not yield any results. Several variables may impact the level of diversity present in an isolate's population. In the context of clinical isolates, there is little control over when the sample is collected with regards to disease progression. An infectious population that is sampled immediately after symptoms manifest may for example vary in its diversity, when compared to a population that is sampled after many days. Furthermore, clinical samples are provided on cryopreserved beads - the capture and subsequent re-culturing provides additional points at which a bottleneck in actual representation of diversity may occur. If these factors that are unique to the clinical datasets did have an impact on the observed diversity, there would likely be a difference in the number of non-variant containing poultry isolates. In reality this is not the case, suggesting that the prevalence of invariable populations amongst *Campylobacter* isolates in general is conserved.

5.5.4. Future Work

The datasets discussed within this chapter provide a novel insight into the diversity of clinical and poultry *Campylobacter spp.* isolates. However, it is important to consider the differences between clinical isolates, poultry isolates that may generate a clinical infection, and the ancestral populations they both may originate from. Figure 5.22 (below) follows a hypothetical *Campylobacter spp.* isolate as it moves from a chicken host to its eventual recovery from a human clinical infection. In the context of this figure we have characterised the diversity of clinical *Campylobacter* populations (D), and produced a preliminary overview of the diversity of isolates from retail poultry (C). However, both populations may have undergone significant changes from the original ancestral isolate (A). Whilst tracking an individual population of *Campylobacter spp.* from a chicken host to a clinical infection is unrealistic, it does highlight additional opportunities for future study. For example, it would be interesting to apply the deep sequencing methodology to additional non-clinical sources (*e.g.* live poultry). This data would allow comparison between the variable loci observed in the clinical and retail poultry datasets, and aid in delineating alleles that are specific to human infection, specific to the poultry host, and perhaps loci that are persistently variable across all hosts.

The clinical datasets show several instances of loci undergoing selective pressure within the human host environment. However, it is not known as to whether these mutations are beneficial, deleterious, or are a result of co-dominant haplotypes, as the clinical isolates represent a single patient sample at a single time point. To investigate this concern, it would be useful to sequence isolates obtained from a single patient, across multiple time points throughout the course of infection. If polymorphisms are seen at increasing or decreasing frequencies it would suggest that variation in this locus is a result of positive or purifying selection. Alternatively, if frequencies are found to occur at stable intermediate levels, it would suggest that multiple lineages are

concurrently present. This data would also aid in the resolution of host specific variation, as outlined above.

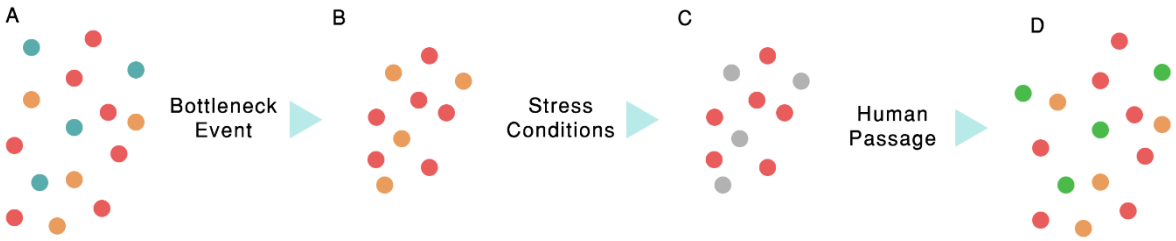

Figure 5.22 – A representation of how haplotypes amongst a hypothetical *Campylobacter spp.* population may change as it moves from a chicken host to causing human infection

A – The starting population inside a chicken host**.**
The starting population may contain several haplotypes. In this example, the starting population contains colonies representing three haplotypes, denoted by their colours (Orange, Red and Blue). Sequencing this population would recover all haplotypes present at this time.
B – The population during slaughter.
Following a bottleneck event (*i.e.* a reduction in the population), the number of haplotypes that are represented amongst the population may be reduced (*e.g.* during carcass rinsing). In this example, the Blue haplotype is no longer present amongst the population – sequencing this population would only detect the Red and Orange haplotypes.
C – The population on retail poultry meat.
If exposed to stress conditions, some colonies may form viable but non culturable cells (grey). In this scenario, the VNC cells originated from a single haplotype (Orange). These colonies would not be analysed by the deep sequencing methodology, which relies on fresh cultures from which DNA is prepared. Sampling at this stage would only reveal the Red haplotype.
D – The population recovered from a clinical infection.
During infection, the VNC cells are resuscitated and now produce culturable forms, restoring the Orange haplotype. During the infection, some cells undergo variation, forming a new haplotype, Green. Deep sequence analysis at this stage would reveal all haplotypes present.

Chapter Six: Conclusion

Campylobacter spp. is the leading cause of bacterial mediated enteritis in the developed world, and yet the genetic epidemiology of campylobacteriosis is not conclusively defined. Existing studies suggest that the majority of cases are sporadic in nature, with no case linkages demonstrated in routine clinical datasets. As an organism, *Campylobacter spp*. undergo a vast amount of genomic variation, and as such it was unclear as to whether the lack of observed case linkage can be truly defined using the assembly based comparative techniques (e.g. wgMLST) that many existing studies employ.

To investigate this further, the work presented in Chapter Three applied a read mapping approach to clinical *Campylobacter spp*. isolates, collected from a single health trust site in Nottinghamshire during 2014. The results from this work presented many novel findings.

Firstly, several instances of case linkage were apparent amongst the dataset, which represents the first description of case linkage between routine clinical isolates. The lack of reported case linkage in previous studies is likely a result of the notable genetic diversity that *Campylobacter spp*. exhibits, as well as the limited sensitivity of the comparative techniques (i.e. wgMLST). In the context of several studies discussed (*i.e.* non-specialised clinicians), utilising this methodology is entirely appropriate, and preferential to the far more resource intensive read-mapping approach. However, where possible (particularly in research environments) it would be advisable to utilise read mapping methods to provide an objective, qualitative measure of true genetic homology. The results presented within this chapter act as a proof of concept, and provide a greater resolving power of both true case linkage, and greater discrimination between false-positives.

This work also identified a period of enhanced incidence during February, 2014. Upon further inspection, this peak in incidence was also

apparent in national data from 2008-2011, which suggests that this effect is not limited to the Nottinghamshire area. This period of increased incidence was also found to contain a higher proportion of linked cases. However, it is not clear as to whether the increase in observed case linkage is genuine, or potentially a result of the increase in percentage of isolates sequenced. In hindsight, it would have been prudent to also sample a week-long tranche of isolates that did not exhibit unexpected incidence to act as a negative control, which would address this concern.

Additionally, the linking of social events to the increase in incidence is purely circumstantial, and it is not clear as to whether patients engaged in any of the described risk factors that may have increased their likelihood of contracting campylobacteriosis. To address this, it would have been useful to obtain additional data from the sampled patients (e.g. questionnaire data).

Due to ethical limitations, this was not possible. The addition of this information would aid in addressing this, as well as many other questions. Campylobacteriosis is a severely underreported disease, and as such a greater incidence of same-source infections may occur at a transient level in a patient's social network. Questionnaire data could reveal whether other friends or family members also fell ill in a capacity that did not provoke the need for medical aid, revealing additional incidence of sporadic case linkage. Additional risk factors could also be investigated with this information (*e.g.* recent foreign travel), and as such future work may consider obtaining this information.

Following the genomic analysis of isolate pairs, further information detailing patient demographics was made available to the study. This revealed that the most genetically homologous isolate pair was in fact isolated from the same patient. Importantly, this isolate pair was not entirely homologous. In fact, over the course of a single day, genetic

variation was apparent. This generated further questions as to whether *Campylobacter spp*. may undergo intra-population variation within a single host.

To investigate this, a subset of clinical isolates was analysed using a deep-sequencing methodology, which greatly increases the coverage depth of sequencing reads and increases the ability to resolve minor allele variation. Initially, this was conducted on the most abundant clinical complex of *C. jejuni* isolates: ST-21, revealing a surprising amount of genetic diversity amongst single clinical populations. This diversity is even more impressive when considering the time scales of typical *Campylobacter spp*. infections, with sampled isolates representing organisms that have been in contact with the human host for ~1-2 weeks.

The deep sequencing methodology uses the *de novo* assembly of each isolate's read data as a reference to which the read data is again mapped against. This is of particular benefit in an organism such as *Campylobacter* that exhibits a vast amount of genetic variation, and as such limits the usability of common reference sequences. However, relying on *de novo* assemblies does present its own issues. Firstly, any areas of the genome that are not appropriately resolved will result in a break of contiguous sequence. This may interrupt CDS, and lead to areas of the genome for which no sequence is defined, preventing alignment. The access and commonality of long-read sequencing has undergone a drastic change since the inception of this study, with the release and constant revisions of technologies such as Oxford Nanopore Technology's MinION device. In future studies, it would be preferable to allocate additional resources to the generation of isolate reference sequences, which would ideally be resolved to a single, high quality contig through a combination of short and long read data. This would also allow an additional layer of verification of called minor allele

variants via mapping of long reads to the generated assembly, and aid in the detection of variants in repetitive sequence elements.

Campylobacter spp. contain several instances of homopolymeric tracts, which are noted sources of genetic diversity. Across all deep-sequenced samples, only four ST-21 isolates revealed variation within these tracts. As discussed in Chapter 4, additional variation may be present amongst the sampled populations, however the strict quality thresholds employed by the deep-sequencing analysis pipeline, coupled with the relatively poor quality of repetitive stretch sequencing on Illumina platforms mean these variants may have been excluded from the final dataset. This may be mitigated by utilising long-read sequencing technologies which offer greater sequencing accuracy of repetitive tracts.

The variants that were identified within the deep sequencing datasets are representative of a population at a single time point. As such, it is not clear as to whether these variants are sweeping to fixation. This provides a key area for future research, wherein repeat samples obtained from a single patient would be analysed using the deep sequencing methodology, allowing observation of changes to the minor allele frequency of identified variants. In the genetic epidemiology work presented in Chapter 3, one such instance of same-patient sampling was revealed. Unfortunately, the patient demographic data was provided after the deep sequencing datasets were generated, and as such it was not possible to include these isolates within the sample.

In an expansion to the ST-21 deep sequencing dataset, isolates from additional CC's were also analysed. Whilst this data identified several interesting points of comparison, it is partially limited by the sample size. In future work, additional sequencing of key lineages (*e.g.* ST-828) would aid in delineating the differences in genomic variation between

clinical isolate complexes. This limitation is also apparent in the poultry dataset.

To detect whether the level of diversity that was observed in single clinical populations of *Campylobacter,* isolates obtained from fresh retail chicken samples were also analysed. This revealed a notable decrease in diversity compared to clinical samples, suggesting that variation is driven within the human host. The reliance on retail chicken isolates was necessary, as an appropriate pre-infectious sample was not available. It would be preferable to analyse an isolate population before and after human passage – future work may consider applying the deep sequence methodology to a confirmed point-source outbreak.

The majority of loci identified as undergoing variation in clinical isolates encode proteins related to chemotaxis functions. Furthermore, many of these variants were found to be contained within the protein's Methyl-accepting signalling domains. The repeated variation within this domain suggests the selective pressure present in the human host is driving variation at these specific locations, and provides an interesting topic for future study. At this stage, it is unclear as to whether this particular pattern of diversity is exclusive to the human host – an absence of similar variants in the poultry dataset, as well as results from an independent study would suggest this is the case, however additional work is required to draw any conclusions. Future studies may consider sampling isolates from live poultry flocks, as well as additional chicken meat samples, environmental isolates, and isolates obtained from animal passage experiments. This data would allow the separation of loci that undergo variation as a result of novel colonisation in general, colonisation of the human host specifically, and loci that may be variable across *Campylobacter* populations in general.

In summary, this work presents several novel findings: Sporadic case linkage can be observed with enhanced comparative techniques, a third peak in incidence is observable across national data when appropriately normalised, and *Campylobacter spp*. exhibits clear patterns of intra-population diversity. This body of work acts as a key foundation for further investigation, with a focus on expansion of the deep sequencing methodology to include additional isolate types and sources, and has important implications for clinical investigation of related cases.

Acknowledgements

Dr. Alan McNally – my first supervisor, a wonderful mentor and great scientist. Your guidance and support* throughout my PhD has been outstanding, I've learned a lot, I've seen you rub a scotch bonnet where no sane man should, and most of all I have thoroughly enjoyed my time working in the McNally lab.

Dr. Georgina Manning – my second supervisor, you are a large part of the reason I chose to pursue a PhD, following an immensely enjoyable undergraduate research project. You are an absolute asset to higher education, and a paragon of professionalism. Thank you for everything you have contributed over the past few years - it would not have been possible without you.

Dr. Benjamin Dickins – my independent assessor, you are still the most intelligent man I have ever met, and I have a feeling you will hold that title for a long time. Your feedback and guidance throughout the various stages of my project's progress are all greatly appreciated.

Mathew Diggle, Jim Turton, Vicki Fleming and all of the microbiology technical staff at Queen's Medical Centre – for providing me access to so many *Campylobacters*, filling my many requests, and for all the important work your laboratory does, thank you.

* - both in academic and alcoholic terms

References

Acheson, D. & Allos, B.M. 2001. *Campylobacter jejuni* Infections: Update on Emerging Issues and Trends. *Clinical Infectious Diseases*, 32, 1201–1206, 10.1086/319760.

Açik, M.N. & Çetinkaya, B. 2006. Heterogeneity of *Campylobacter jejuni* and *Campylobacter coli* strains from healthy sheep. *Veterinary Microbiology*, 115, 370–375, 10.1016/j.vetmic.2006.02.014.

Agashe, D., Sane, M., Phalnikar, K., Diwan, G.D., Habibullah, A., Martinez-Gomez, N.C., Sahasrabuddhe, V. 2016. Large-Effect Beneficial Synonymous Mutations Mediate Rapid and Parallel Adaptation in a Bacterium. *Molecular Biology and Evolution*, 33, 1542–1553, 10.1093/molbev/msw035.

Aidley, J. & Bayliss, C.D. 2014. Repetitive DNA: a major source of genetic diversity in *Campylobacter* populations? *In Campylobacter Ecology and Evolution*. Portland: Caister Academic Press.

Alexander, R.P., Lowenthal, A.C., Harshey, R.M. & Ottemann, K.M. 2010. CheV: CheW-like coupling proteins at the core of the chemotaxis signaling network. *Trends in Microbiology*, 18, 494–503, 10.1016/j.tim.2010.07.004.

Alkan, C., Sajjadian, S. & Eichler, E.E. 2011. Limitations of next-generation genome sequence assembly. *Nature Methods*, 8, 61–65, 10.1038/nmeth.1527.

Allerberger, F., Al-Jazrawi, N., Kreidl, P., Dierich, M.P., Feierl, G., Hein, I. & Wagner, M. 2003. Barbecued chicken causing a multi-state outbreak of *Campylobacter jejuni* enteritis. *Infection*, 31, 19–23, 10.1007/s15010-002-3088-8.

Altmann, A., Weber, P., Bader, D., Preuß, M., Binder, E.B. & Müller-Myhsok, B. 2012. A beginners guide to SNP calling from high-Throughput DNA-sequencing data. *Human Genetics*, 131, 1541–1554, 10.1007/s00439-012-1213-z.

Ang, C.W., Noordzij, P.G., De Klerk, M.A., Endtz, H.P., Van Doorn, P.A. & Laman, J.D. 2002. Ganglioside mimicry of *Campylobacter jejuni* lipopolysaccharides determines antiganglioside specificity in

rabbits. *Infection and Immunity*, 70, 5081–5085, 10.1128/IAI.70.9.5081-5085.2002.

Ardley, C. & Wright, S. 2010. Travellers' diarrhoea. *Medicine*, 38, 26–29, 10.1016/j.mpmed.2009.09.028.

Ashbolt, N.J. 2004. Microbial contamination of drinking water and disease outcomes in developing regions. *In Toxicology*. 229–238., 10.1016/j.tox.2004.01.030.

Ashgar, S.S.A., Oldfield, N.J., Wooldridge, K.G., Jones, M.A., Irving, G.J., Turner, D.P.J. & Ala'Aldeen, D. 2007. CapA, an autotransporter protein of *Campylobacter jejuni*, mediates association with human epithelial cells and colonization of the chicken gut. *Journal of Bacteriology*, 189, 1856–1865, 10.1128/JB.01427-06.

Bacon, D.J., Alm, R., Burr, D.H., Hu, L., Kopecko, D.J., Ewing, C.P., Trust, T.J. & Guerry, P. 2000. Involvement of a plasmid in virulence of *Campylobacter jejuni* 81-176. *Infection and Immunity*, 68, 4384–4390, 10.1128/IAI.68.8.4384-4390.2000.

Baker, J., Barton, M.D. & Lanser, J. 1999. *Campylobacter* species in cats and dogs in South Australia. *Australian veterinary journal*, 77, 662–666, DOI 10.1111/j.1751-0813.1999.tb13159.x.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19, 455–477, 10.1089/cmb.2012.0021.

Baqar, S., Applebee, L.A., Gilliland, T.C., Lee, L.H., Porter, C.K. & Guerry, P. 2008. Immunogenicity and protective efficacy of recombinant *Campylobacter jejuni* flagellum-secreted proteins in mice. *Infection and Immunity*, 76, 3170–3175, 10.1128/IAI.00076-08.

Bayliss, C.D., Bidmos, F.A., Anjum, A., Manchev, V.T., Richards, R.L., Grossier, J.P., Wooldridge, K.G. 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, 5876–5889, 10.1093/nar/gks246.

Becker, D.J. & Lowe, J.B. 2003. Fucose: Biosynthesis and biological function in mammals. *Glycobiology*, 13, 41R–53R, 10.1093/glycob/cwg054.

Beeby, M., Ribardo, D.A., Brennan, C.A., Ruby, E.G., Jensen, G.J. & Hendrixson, D.R. 2016. Diverse high-torque bacterial flagellar motors assemble wider stator rings using a conserved protein scaffold. *Proceedings of the National Academy of Sciences*, 113, E1917–E1926, 10.1073/pnas.1518952113.

Berndtson, E., Danielsson-Tham, M.L. & Engvall, A. 1996. *Campylobacter* incidence on a chicken farm and the spread of *Campylobacter* during the slaughter process. *International Journal of Food Microbiology*, 32, 35–47, 10.1016/0168-1605(96)01102-6.

Bhaduri, S. & Cottrell, B. 2004. Survival of cold-stressed *Campylobacter jejuni* on ground chicken and chicken skin during frozen storage. *Applied and Environmental Microbiology*, 70, 7103–7109, 10.1128/AEM.70.12.7103-7109.2004.

Bi, P., Cameron, A.S., Zhang, Y. & Parton, K.A. 2008. Weather and notified *Campylobacter* infections in temperate and sub-tropical regions of Australia: An ecological study. *Journal of Infection*, 57, 317–323, 10.1016/j.jinf.2008.08.004.

Black, R.E., Levine, M.M., Clements, M. Lou, Hughes, T.P. & Blaser, M.J. 1988. Experimental *Campylobacter jejuni* infection in humans. *Journal of Infectious Diseases*, 157, 472–479, 10.1093/infdis/157.3.472.

Blaser, M.J., Berkowitz, I.D., LaForce, F.M., Cravens, J., Reller, L.B. & Wang, W.L. 1979. *Campylobacter* enteritis: Clinical and epidemiologic features. *Annals of Internal Medicine*, 91, 179–185, 10.7326/0003-4819-91-2-179.

Blaser, M.J., Wells, J.G., Feldman, R.A., Pollard, R.A. & Allen, J.R. 1983. *Campylobacter* enteritis in the United States. A multicenter study. *Annals of Internal Medicine*, 98, 360–365, 10.7326/0003-4819-98-3-360.

Blaser, M.J., Black, R.E., Duncan, D.J. & Amer, J. 1985. *Campylobacter jejuni*-specific serum antibodies are elevated in healthy Bangladeshi children. *Journal of clinical microbiology*, 21, 164–167.

Blaser, M.J. 1997. Epidemiologic and clinical features of *Campylobacter jejuni* infections. *J Infect.Dis*, 176 Suppl, S103–S105.

Blaser, MJ; Engberg, J. 2008. Clinical aspects of *Campylobacter jejuni* and *Campylobacter coli* infections. *In Campylobacter. Vol.3*. American Society of Microbiology, 99–121., 10.1128/9781555815554.ch6.

Boehm, M., Lind, J., Backert, S. & Tegtmeyer, N. 2015. *Campylobacter jejuni* serine protease HtrA plays an important role in heat tolerance, oxygen resistance, host cell adhesion, invasion, and transmigration. *European Journal of Microbiology and Immunology*, 5, 68–80, 10.1556/EuJMI-D-15-00003.

Bokkenheuser, V.D., Richardson, N.J., Bryner, J.H., Roux, D.J., Schutte, A.B., Koornhof, H.J., Freiman, I. & Hartman, E. 1979. Detection of enteric campylobacteriosis in children. *Journal of Clinical Microbiology*, 9, 227–232.

Broczyk, A., Thompson, S., Smith, D. & Lior, H. 1987. Water-Borne Outbreak of *Campylobacter* Gastroenteritis. *The Lancet*, 329, 164–165, 10.1016/S0140-6736(87)92003-4.

Burnens, A., Stucki, U., Nicolet, J. & Frey, J. 1995. Identification and characterization of an immunogenic outer membrane protein of *Campylobacter jejuni*. *Journal of Clinical Microbiology*, 33, 2826–2832.

Butzler, J.-P. 2004. *Campylobacter*, from obscurity to celebrity. *Clin Microbiol Infect*, 10, 868–876, 10.1111/j.1469-0691.2004.00983.x.

Cawthraw, S. a, Wassenaar, T.M., Ayling, R. & Newell, D.G. 1996. Increased colonization potential of *Campylobacter jejuni* strain 81116 after passage through chickens and its implication on the rate of transmission within flocks. *Epidemiology and infection*, 117, 213–215, 10.1017/S0950268800001333.

Cean, A., Nicolae, C., Lavinia, S., Gabi, D., Sărăndan, M., Alina, G., Călin, J. & Dan, D. 2013. Isolation of *Campylobacter jejuni* from cloaca and cecum content of chicken broilers bred in intensive systems in the Western part of Romani. *Scientific Papers Animal Science and Biotechnologies*, 46, 93–97.

Champion, O. 2005. Comparative phylogenomics of the food-borne pathogen *Campylobacter jejuni* reveals genetic markers predictive of infection source. *Proceedings of the …*, 102, 16043–16048, 10.1073/pnas.0503252102.

Chen, J., Sun, X.-T., Zeng, Z. & Yu, Y.-Y. 2011. *Campylobacter* enteritis in adult patients with acute diarrhea from 2005 to 2009 in Beijing, China. *Chinese medical journal*, 124, 1508–1512.

Clark, C.G. & Ng, L.-K. 2008. Sequence variability of *Campylobacter* temperate bacteriophages. *BMC Microbiology*, 8, 49, 10.1186/1471-2180-8-49.

Cody, A.J., Maiden, M.J.C. & Dingle, K.E. 2009. Genetic diversity and stability of the porA allele as a genetic marker in human *Campylobacter* infection. *Microbiology*, 155, 4145–4154, 10.1099/mic.0.031047-0.

Cody, A.J., McCarthy, N.M., Wimalarathna, H.L., Colles, F.M., Clark, L., Bowler, I.C.J.W., Maiden, M.C.J. & Dingle, K.E. 2012. A longitudinal 6-year study of the molecular epidemiology of clinical *Campylobacter* isolates in Oxfordshire, United Kingdom. *Journal of Clinical Microbiology*, 50, 3193–3201, 10.1128/JCM.01086-12.

Cody, A.J., McCarthy, N.D., Van Rensburg, M.J., Isinkaye, T., Bentley, S.D., Parkhill, J., Dingle, K.E., Bowler, I.C.J.W., Jolley, K.A. & Maiden, M.C.J. 2013. Real-time genomic epidemiological evaluation of human *Campylobacter* isolates by use of whole-genome multilocus sequence typing. *Journal of Clinical Microbiology*, 51, 2526–2534, 10.1128/JCM.00066-13.

Coker, A.O., Isokpehi, R.D., Thomas, B.N., Amisu, K.O. & Larry Obi, C. 2002. Human campylobacteriosis in developing countries. *Emerging*

Infectious Diseases, 8, 237–243, 10.3201/eid0803.010233.

Colles, F.M., Jones, K., Harding, R.M. & Maiden, M.C.J. 2003. Genetic Diversity of *Campylobacter jejuni* Isolates from Farm Animals and the Farm Environment. *Applied and Environmental Microbiology*, 69, 7409–7413, 10.1128/AEM.69.12.7409-7413.2003.

Corry, J.E. & Atabay, H.I. 2001. Poultry as a source of *Campylobacter* and related organisms. *Symposium series (Society for Applied Microbiology)*, 96S–114S, 10.1046/j.1365-2672.2001.01358.x.

Corry, J.E.L., Atabay, H.I., Forsythe, S.J. & Mansfield, L.P. 2003. Culture media for the isolation of *Campylobacter*s, helicobacters and arcobacters. *Progress in Industrial Microbiology*, 37, 271–316, 10.1016/S0079-6352(03)80021-8.

Crawshaw, T.R., Chanter, J.I., Young, S.C.L., Cawthraw, S., Whatmore, A.M., Koylass, M.S., Vidal, A.B., Salguero, F.J. & Irvine, R.M. 2015. Isolation of a novel thermophilic *Campylobacter* from cases of spotty liver disease in laying hens and experimental reproduction of infection and microscopic pathology. *Veterinary microbiology*, 179, 315–321, 10.1016/j.vetmic.2015.06.008.

Cunningham, C., Lee, C.H., Passarino, G., Burlo, P., Ciccone, G., Jin, O., Sole, M. 2003. Myocarditis related to *Campylobacter jejuni*infection: A case report. *BMC Infectious Diseases*, 3, 16, 10.1186/1471-2334-3-16.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E. 2011. The variant call format and VCFtools. *Bioinformatics*, 27, 2156–2158, 10.1093/bioinformatics/btr330.

Davies, J.C. 2002. Pseudomonas aeruginosa in cystic fibrosis: Pathogenesis and persistence. *Paediatric Respiratory Reviews*, 3, 128–134, 10.1016/S1526-0550(02)00003-3.

Davis, L. & DiRita, V. 2008. Growth and laboratory maintenance of *Campylobacter jejuni*. *Current Protocols in Microbiology*, 10.1002/9780471729259.mc08a01s10.

Dearlove, B.L., Cody, A.J., Pascoe, B., Méric, G., Wilson, D.J., Sheppard, S.K. & Daniel, J. 2015. Rapid host switching in generalist

Campylobacter strains erodes the signal for tracing human infections. *The ISME journal*, 10, 1–9, 10.1038/ismej.2015.149.

De Zoete, M.R., Keestra, A.M., Wagenaar, J.A. & Van Putten, J.P.M. 2010. Reconstitution of a functional toll-like receptor 5 binding site in *Campylobacter jejuni* flagellin. *Journal of Biological Chemistry*, 285, 12149–12158, 10.1074/jbc.M109.070227.

DEFRA. 2015. Zoonoses Report UK 2013. *Public Health England*, 1–77.

DEFRA & PHE. 2015. Zoonoses Summary Report - UK 2014.

Dimachkie, M.M. & Barohn, R.J. 2013. Guillain-Barré syndrome and variants. *Neurologic Clinics*, 31, 491–510, 10.1016/j.ncl.2013.01.005.

Dingle, K.E., Colles, F.M., Wareing, D.R.A.A., Ure, R., Fox, A.J., Bolton, F.E., Bootsma, H.J., Willems, R.J.L., Urwin, R. & Maiden, M.C.J. 2001. Multilocus Sequence Typing System for *Campylobacter jejuni*. *Journal of clinical microbiology*, 39, 14–23, 10.1128/JCM.39.1.14.

Dingle, K.E., Colles, F.M., Ure, R., Wagenaar, J.A., Duim, B., Bolton, F.J., Fox, A.J., Wareing, D.R.A. & Maiden, M.C.J. 2002. Molecular Characterization of *Campylobacter jejuni* Clones: A Basis for Epidemiologic Investigation. *Emerging Infectious Diseases*, 8, 949–955, 10.3201/eid0809.020122.

Dong, H., Nilsson, L. & Kurland, C.G. 1996. Co-variation of tRNA abundance and codon usage in Escherichia *coli* at different growth rates. *Journal of molecular biology*, 260, 649–663, 10.1006/jmbi.1996.0428.

Doorduyn, Y., Van Den Brandhof, W.E., Van Duynhoven, Y.T.H.P., Breukink, B.J., Wagenaar, J.A. & Van Pelt, W. 2010. Risk factors for indigenous *Campylobacter jejuni* and *Campylobacter coli* infections in The Netherlands: a case-control study. *Epidemiology and infection*, 138, 1391–1404, 10.1017/S095026881000052X.

Eberhart-Phillips, J., Walker, N., Garrett, N., Bell, D., Sinclair, D., Rainger, W. & Bates, M. 1997. Campylobacteriosis in New Zealand: results of a case-control study. *Journal of epidemiology and community health*, 51, 686–691, 10.1136/jech.51.6.686.

EFSA. 2011a. Scientific Opinion on *Campylobacter* in broiler meat production: control options and performance objectives and/or targets at different stages of the food chain. *EFSA Journal*, 9, 1–141, 10.2903/j.efsa.2011.2105.

EFSA. 2014. The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2014 European Food Safety Authority European Centre for Disease Prevention and Control. *EFSA Journal*, 13, 10.2903/j.efsa.2015.4329.

EFSA. 2011b. Tracing seeds, in particular fenugreek (Trigonella foenum-graecum) seeds, in relation to the Shiga toxin-producing E. *coli* (STEC) O104:H4. 2011, http://www.efsa.europa.eu/en/supporting/pub/176e.htm, Outbreaks in Germany and France,. *EFSA Supporting Publications*, 8, 23, 10.2903/sp.efsa.2011.EN-176.

Ekdahl, K., Normann, B. & Andersson, Y. 2005. Could flies explain the elusive epidemiology of campylobacteriosis? *BMC Infectious Diseases*, 5, 11, 10.1186/1471-2334-5-11.

Etoh, Y., Dewhirst, F.E., Paster, B.J., Yamamoto, A. & Goto, N. 1993. *Campylobacter* showae sp. nov., isolated from the human oral cavity. *International journal of systematic bacteriology*, 43, 631–639, 10.1099/00207713-43-4-631.

Evans, S.J. & Sayers, A.R. 2000. A longitudinal study of *Campylobacter* infection of broiler flocks in Great Britain. *Preventive Veterinary Medicine*, 46, 209–223, 10.1016/S0167-5877(00)00143-4.

Fahey, T., Morgan, D., Gunneburg, C., Adak, G.K., Majid, F. & Kaczmarski, E. 1995. An outbreak of *Campylobacter jejuni* enteritis associated with failed milk pasteurisation. *Journal of Infection*, 31, 137–143, 10.1016/S0163-4453(95)92160-5.

FAO. 2016. *FAOSTAT: FAO Statistical Database*.

Fernández-Cruz, A., Muñoz, P., Mohedano, R., Valerio, M., Marín, M., Alcalá, L., Rodriguez-Créixems, M., Cercenado, E. & Bouza, E. 2010. *Campylobacter* Bacteremia. *Medicine*, 89, 319–330,

10.1097/MD.0b013e3181f2638d.

Fernández, H., Vera, F., Villanueva, M.P. & García, A. 2008. Occurrence of *Campylobacter* species in healthy well-nourished and malnourished children. *Brazilian Journal of Microbiology*, 39, 56–58, 10.1590/S1517-83822008000100013.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C. 2016. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research*, 44, D279–D285, 10.1093/nar/gkv1344.

Fitzgerald, C. & Nachamkin, I. 2011. *Campylobacter* and Arcobacter. *Manual of Clinical Microbiology, 10th Edition*, 885–899, 10.1128/9781555816728.ch53.

Flanagan, R.C., Neal-McKinney, J.M., Dhillon, A.S., Miller, W.G. & Konkel, M.E. 2009. Examination of *Campylobacter jejuni* putative adhesins leads to the identification of a new protein, designated FlpA, required for chicken colonization. *Infection and Immunity*, 77, 2399–2407, 10.1128/IAI.01266-08.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J. 1995. Whole-Genome Random Sequencing and Assembly of Haemophilus-Influenzae Rd. *Science*, 269, 496–512, 10.1126/science.7542800.

Ford, C.B., Lin, P.L., Chase, M.R., Shah, R.R., Iartchouk, O., Galagan, J., Mohaideen, N. 2011. Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. *Nature Genetics*, 43, 482–486, 10.1038/ng.811.

Fouts, D.E., Mongodin, E.F., Mandrell, R.E., Miller, W.G., Rasko, D.A., Ravel, J., Brinkac, L.M. 2005. Major structural differences and novel potential virulence mechanisms from the genomes of multiple *Campylobacter* species. *PLoS Biology*, 3, 11168, 10.1371/journal.pbio.0030015.

French, N.P., Midwinter, A., Holland, B., Collins-Emerson, J., Pattison, R., Colles, F. & Carter, P. 2009. Molecular epidemiology of

Campylobacter jejuni isolates from wild-bird fecal material in children's playgrounds. *Applied and Environmental Microbiology*, 75, 779–783, 10.1128/AEM.01979-08.

Fricke, W.F. & Rasko, D.A. 2013. Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nature Reviews Genetics*, 15, 49–55, 10.1038/nrg3624.

Frost, J. a. 2001. Current epidemiological issues in human campylobacteriosis. *Symposium series (Society for Applied Microbiology)*, 90, 85S–95S, 10.1046/j.1365-2672.2001.01357.x.

FSA & BBSRC. 2013. Summary of the 2013 *Campylobacter* Strategy Workshop.

FSA & DEFRA. 2010. the Joint Government and Industry Target To Reduce *Campylobacter* in Uk Produced Chickens By 2015. 1–30.

FSA. 2016. *Campylobacter* contamination in fresh whole chilled UK-produced chickens at retail: January – March 2016.

Garénaux, A. 2008. *Better understanding the Campylobacter conundrum*. Nova Science Publishers, 151 pp.

Garmory, H.S. & Titball, R.W. 2004. ATP-binding cassette transporters are targets for the development of antibacterial vaccines and therapies. *Infection and Immunity*, 72, 6757–6763, 10.1128/IAI.72.12.6757-6763.2004.

Gebhart, C.J., Edmonds, P., Ward, G.E., Kurtz, H.J. & Brenner, D.J. 1985. "*Campylobacter* hyointestinalis" sp. nov.: a new species of *Campylobacter* found in the intestines of pigs and other animals. *Journal of clinical microbiology*, 21, 715–720.

Gegner, J.A., Graham, D.R., Roth, A.F. & Dahlquist, F.W. 1992. Assembly of an MCP receptor, CheW, and kinase CheA complex in the bacterial chemotaxis signal transduction pathway. *Cell*, 70, 975–982, 10.1016/0092-8674(92)90247-A.

Gibbens, J.C., Pascoe, S.J.S., Evans, S.J., Davies, R.H. & Sayers, A.R. 2001. A trial of biosecurity as a means to control *Campylobacter* infection of broiler chickens. *Preventive Veterinary Medicine*, 48, 85–99,

10.1016/S0167-5877(00)00189-6.

Gilbert M, Parker CT & Moran, A. 2008. *Campylobacter jejuni* Lipooligosaccharides: Structures and Biosynthesis. *Campylobacter book*, Book chapt, 493–504, 10.1128/9781555815554.ch27.

Gilbert, M.J., Kik, M., Miller, W.G., Duim, B. & Wagenaar, J.A. 2015. *Campylobacter* iguaniorum sp. nov., isolated from reptiles. *International Journal of Systematic and Evolutionary Microbiology*, 65, 975–982, 10.1099/ijs.0.000048.

Gillespie, I.A., O'Brien, S.J., Frost, J.A., Adak, G.K., Horby, P., Swan, A. V., Painter, M.J. & Neal, K.R. 2002. A case-case comparison of *Campylobacter coli* and *Campylobacter jejuni* infection: A tool for generating hypotheses. *Emerging Infectious Diseases*, 8, 937–942, 10.3201/eid0809.010187.

Gillespie, I.A., O'Brien, S.J., Frost, J.A., Tam, C., Tompkins, D., Neal, K.R., Syed, Q. & Farthing, M.J.G. 2006. Investigating vomiting and/or bloody diarrhoea in *Campylobacter jejuni* infection. *Journal of Medical Microbiology*, 55, 741–746, 10.1099/jmm.0.46422-0.

Gillespie, I.A., O'Brien, S.J. & Bolton, F.E. 2009. Age patterns of persons with campylobacteriosis, England and Wales, 1990-2007. *Emerging Infectious Diseases*, 15, 2046–2048, 10.3201/eid1512.090280.

Godschalk, P.C.R., Heikema, A.P., Gilbert, M., Komagamine, T., Wim Ang, C., Glerum, J., Brochu, D . 2004. The crucial role of *Campylobacter jejuni* genes in anti-ganglioside antibody induction in Guillain-Barré syndrome. *Journal of Clinical Investigation*, 114, 1659–1665, 10.1172/JCI200415707.

Golden, N.J. & Acheson, D.W.K. 2002. Identification of motility and autoagglutination *Campylobacter jejuni* mutants by random transposon mutagenesis. *Infection and immunity*, 70, 1761–1771, 10.1128/IAI.70.4.1761-1771.2002.

Grad, Y.H., Lipsitch, M., Feldgarden, M., Arachchi, H.M., Cerqueira, G.C., FitzGerald, M., Godfrey, P. 2012. Genomic epidemiology of the Escherichia *coli* O104:H4 outbreaks in Europe, 2011. *Proceedings of*

the National Academy of Sciences, 109, 3065–3070, 10.1073/pnas.1121491109.

Gras, L.M., Smid, J.H., Wagenaar, J. a, Koene, M.G.J., Havelaar, a H., Friesema, I.H.M., French, N.P. 2013. Increased risk for *Campylobacter jejuni* and C. *coli* infection of pet origin in dog owners and evidence for genetic association between strains causing infection in humans and their pets. *Epidemiology and infection*, 141, 2526–2535, 10.1017/S0950268813000356.

Guerrant, R.L., Lahita, R.G., Winn, W.C. & Roberts, R.B. 1978. Campylobacteriosis in man: Pathogenic mechanisms and review of 91 bloodstream infections. *The American Journal of Medicine*, 65, 584–592, 10.1016/0002-9343(78)90845-8.

Guerry, P. 2007. *Campylobacter* flagella: not just for motility. *Trends in Microbiology*, 15, 456–461, 10.1016/j.tim.2007.09.006.

Gundogdu, O., Bentley, S.D., Holden, M.T., Parkhill, J., Dorrell, N. & Wren, B.W. 2007. Re-annotation and re-analysis of the *Campylobacter jejuni* NCTC11168 genome sequence. *BMC Genomics*, 8, 162, 10.1186/1471-2164-8-162.

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

Gutierrez, M., Brisse, S., Brosch, R., Fabre, M., Omaïs, B., Marmiesse, M., Supply, P. & Vincent, V. 2005. Ancient origin and gene mosaicism of the progenitor of Mycobacterium tuberculosis. *PLoS Pathogens*, 1, 0055–0061, 10.1371/journal.ppat.0010005.

Havelaar, A.H., Haagsma, J.A., Mangen, M.J.J., Kemmeren, J.M., Verhoef, L.P.B., Vijgen, S.M.C., Wilson, M. 2012. Disease burden of foodborne pathogens in the Netherlands, 2009. *International Journal of Food Microbiology*, 156, 231–238, 10.1016/j.ijfoodmicro.2012.03.029.

Hayashi, T. 2001. Complete Genome Sequence of Enterohemorrhagic Eschelichia *coli* O157:H7 and Genomic Comparison with a

Laboratory Strain K-12. *DNA Research*, 8, 11–22, 10.1093/dnares/8.1.11.

Henderson, I.R., Owen, P. & Nataro, J.P. 1999. Molecular switches--the ON and OFF of bacterial phase variation. *Molecular microbiology*, 33, 919–932, 10.1046/j.1365-2958.1999.01555.x.

Henderson, I.R., Navarro-Garcia, F., Desvaux, M., Fernandez, R.C. & Ala'Aldeen, D. 2004. Type V Protein Secretion Pathway: the Autotransporter Story. *Microbiology and Molecular Biology Reviews*, 68, 692–744, 10.1128/MMBR.68.4.692-744.2004.

Hendrix, R.W., Lawrence, J.G., Hatfull, G.F. & Casjens, S. 2000. The origins and ongoing evolution of viruses. *Trends in Microbiology*, 8, 504–508, 10.1016/S0966-842X(00)01863-1.

Hendrixson, D.R., Akerley, B.J. & DiRita, V.J. 2001. Transposon mutagenesis of *Campylobacter jejuni* identifies a bipartite energy taxis system required for motility. *Molecular Microbiology*, 40, 214–224, 10.1046/j.1365-2958.2001.02376.x.

Hendrixson, D.R. & DiRita, V.J. 2004. Identification of *Campylobacter jejuni* genes involved in commensal colonization of the chick gastrointestinal tract. *Molecular Microbiology*, 52, 471–484, 10.1111/j.1365-2958.2004.03988.x.

Hermans, D., Van Deun, K., Martel, A., Van Immerseel, F., Messens, W., Heyndrickx, M., Haesebrouck, F. & Pasmans, F. 2011. Colonization factors of *Campylobacter jejuni* in the chicken gut. *Veterinary Research*, 42, 82, 10.1186/1297-9716-42-82.

Hermans, D., Pasmans, F., Messens, W., Martel, A., Van Immerseel, F., Rasschaert, G., Heyndrickx, M., Van Deun, K. & Haesebrouck, F. 2012. Poultry as a host for the zoonotic pathogen *Campylobacter jejuni*. *Vector borne and zoonotic diseases (Larchmont, N.Y.)*, 12, 89–98, 10.1089/vbz.2011.0676.

Herring, C.D., Raghunathan, A., Honisch, C., Patel, T., Applebee, M.K., Joyce, A.R., Albert, T.J. 2006. Comparative genome sequencing of Escherichia *coli* allows observation of bacterial evolution on a

laboratory timescale. *Nature Genetics*, 38, 1406–1412, 10.1038/ng1906.

Heuvelink, A.E., van Heerwaarden, C., Zwartkruis-Nahuis, A., Tilburg, J.J.H.C., Bos, M.H., Heilmann, F.G.C., Hofhuis, A., Hoekstra, T. & de Boer, E. 2009. Two outbreaks of campylobacteriosis associated with the consumption of raw cows' milk. *International Journal of Food Microbiology*, 134, 70–74, 10.1016/j.ijfoodmicro.2008.12.026.

Hofreuter, D., Tsai, J., Watson, R.O., Novik, V., Altman, B., Benitez, M., Clark, C. 2006. Unique features of a highly pathogenic *Campylobacter jejuni* strain. *Infection and Immunity*, 74, 4694–4707, 10.1128/IAI.00210-06.

Holt, K.E., Wertheim, H., Zadoks, R.N., Baker, S., Whitehouse, C.A., Dance, D., Jenney, A. 2015. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae* , an urgent threat to public health. *Proceedings of the National Academy of Sciences*, 112, E3574–E3581, 10.1073/pnas.1501049112.

Howard, S.L., Jagannathan, A., Soo, E.C., Hui, J.P.M., Aubry, A.J., Ahmed, I., Karlyshev, A. 2009. *Campylobacter jejuni* glycosylation island important in cell charge, legionaminic acid biosynthesis, and colonization of chickens. *Infection and Immunity*, 77, 2544–2556, 10.1128/IAI.01425-08.

Hu, L. & Kopecko, D.J. 1999. *Campylobacter jejuni* 81-176 associates with microtubules and dynein during invasion of human intestinal cells. *Infection and Immunity*, 67, 4171–4182.

Hu, L., Raybourne, R.B. & Kopecko, D.J. 2005. Ca2+ release from host intracellular stores and related signal transduction during *Campylobacter jejuni* 81-176 internalization into human intestinal cells. *Microbiology*, 151, 3097–3105, 10.1099/mic.0.27866-0.

Hu, L. & Kopecko, J. 2008. Cell Biology of Human Host Cell Entry by *Campylobacter jejuni*. *Campylobacter, Third Edition*, 297–313, 10.1128/9781555815554.ch17.

Huang, S., Sahin, O. & Zhang, Q. 2007. Infection-induced antibodies against the major outer membrane protein of *Campylobacter jejuni* mainly recognize conformational epitopes. *FEMS Microbiology Letters*, 272, 137–143, 10.1111/j.1574-6968.2007.00752.x.

Hugdahl, M.B., Beery, J.T. & Doyle, M.P. 1988. Chemotactic behavior of *Campylobacter jejuni*. *Infection and Immunity*, 56, 1560–1566.

Humphrey, T.J. 1989. An appraisal of the efficacy of pre-enrichment for the isolation of *Campylobacter jejuni* from water and food. *The Journal of applied bacteriology*, 66, 119–126, DOI 10.1111/j.1365-2672.1989.tb02461.x.

Humphrey, T.J., Henley, A. & Lanning, D.G. 1993. The colonization of broiler chickens with *Campylobacter jejuni*: some epidemiological investigations. *Epidemiology and infection*, 110, 601–607, 10.1017/S0950268800051025.

Ikeda, N. & Karlyshev, A. V. 2012. Putative mechanisms and biological role of coccoid form formation in *Campylobacter jejuni*. *European journal of microbiology & immunology*, 2, 41–49, 10.1556/EuJMI.2.2012.1.7.

Ingale, A.G. & Goto, S. 2013. Identification of antigenic epitopes, homology modeling and structural characterization of capsule biosynthesis protein (CapA) from *Campylobacter jejuni*. *Gene Therapy and Molecular Biology*, 15, 74–84.

Inglis, G.D., Kalischuk, L.D. & Busz, H.W. 2003. A survey of *Campylobacter* species shed in faeces of beef cattle using polymerase chain reaction. *Canadian Journal of Microbiology*, 49, 655–661, 10.1139/w03-087.

Inns, T., Foster, K. & Gorton, R. 2010. Cohort study of a campylobacteriosis outbreak associated with chicken liver parfait, United Kingdom, june 2010. *Eurosurveillance*, 15, 1–4.

Inouye, M., Dashnow, H., Raven, L.-A., Schultz, M.B., Pope, B.J., Tomita, T., Zobel, J. & Holt, K.E. 2014. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Medicine*, 6,

90, 10.1186/s13073-014-0090-6.

Iraola, G., Perez, R., Naya, H., Paolicchi, F., Pastor, E., Valenzuela, S., Calleros, L., Velilla, A., Hernandez, M. & Morsella, C. 2014. Genomic evidence for the emergence and evolution of pathogenicity and niche preferences in the genus *Campylobacter*. *Genome Biology and Evolution*, 6, 2392–2405, 10.1093/gbe/evu195.

Janssen, R., Krogfelt, K.A., Cawthraw, S.A., Van Pelt, W., Wagenaar, J.A. & Owen, R.J. 2008. Host-pathogen interactions in *Campylobacter* infections: The host perspective. *Clinical Microbiology Reviews*, 21, 505–518, 10.1128/CMR.00055-07.

Jerome, J.P., Bell, J.A., Plovanich-Jones, A.E., Barrick, J.E., Brown, C.T. & Mansfield, L.S. 2011. Standing genetic variation in contingency loci drives the rapid adaptation of *Campylobacter jejuni* to a novel host Bereswill, S., *ed*. *PLoS ONE*, 6, e16399, 10.1371/journal.pone.0016399.

Jokinen, C.C., Koot, J.M., Carrillo, C.D., Gannon, V.P.J., Jardine, C.M., Mutschall, S.K., Topp, E. & Taboada, E.N. 2012. An enhanced technique combining pre-enrichment and passive filtration increases the isolation efficiency of *Campylobacter jejuni* and *Campylobacter coli* from water and animal fecal samples. *Journal of Microbiological Methods*, 91, 506–513, 10.1016/j.mimet.2012.09.005.

Jolley, K.A. & Maiden, M.C. 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*, 11, 595, 10.1186/1471-2105-11-595.

Jones, K. 2001. *Campylobacter*s in water, sewage and the environment. *Journal of Applied Microbiology*, 90, 68S–79S, 10.1046/j.1365-2672.2001.01355.x.

Joshi, M. 2014. *Dictionary of root words : Greek and Latin roots*, 110 pp.

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

Kapperud, G., Lassen, J., Ostroff, S.M. & Aasen, S. 1992. Clinical features of sporadic *Campylobacter* infections in Norway. *Scandinavian*

journal of infectious diseases, 24, 741–749, Doi 10.3109/00365549209062459.

Kapperud, G., Skjerve, E., Bean, N.H., Ostroff, S.M. & Lassen, J. 1992. Risk factors for sporadic *Campylobacter* infections: Results of a case-control study in southeastern Norway. *Journal of Clinical Microbiology*, 30, 3117–3121.

Kärenlampi, R., Rautelin, H., Schönberg-Norio, D., Paulin, L. & Hänninen, M.L. 2007. Longitudinal study of Finnish *Campylobacter jejuni* and C. *coli* isolates from humans, using multilocus sequence typing, including comparison with epidemiological data and isolates from poultry and cattle. *Applied and Environmental Microbiology*, 73, 148–155, 10.1128/AEM.01488-06.

Kelly, D.J. 2001. The physiology and metabolism of *Campylobacter jejuni* and Helicobacter pylori. *Journal of Applied Microbiology*, 90, 16S–24S, 10.1046/j.1365-2672.2001.01350.x.

Kennemann, L., Didelot, X., Aebischer, T., Kuhn, S., Drescher, B., Droege, M., Reinhardt, R. 2011. Helicobacter pylori genome evolution during human infection. *Proceedings of the National Academy of Sciences*, 108, 5033–5038, 10.1073/pnas.1018444108.

Ketley, J.M. & Konkel, M.E. 2005. *Campylobacter : molecular and cellular biology*.

Key Note. 2015. *Poultry: Market Update*.

Kiehlbauch, J.A., Brenner, D.J., Nicholson, M.A., Baker, C.N., Patton, C.M., Steigerwalt, A.G. & Wachsmuth, I.K. 1991. *Campylobacter* butzleri sp. nov. isolated from humans and animals with diarrheal illness. *Journal of Clinical Microbiology*, 29, 376–385.

Kim, J.S., Artymovich, K.A., Hall, D.F., Smith, E.J., Fulton, R., Bell, J., Dybas, L. 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, 1304–1316, 10.1099/mic.0.057158-0.

Klena, J.D., Parker, C.T., Knibb, K., Claire Ibbitt, J., Devane, P.M.L., Horn, S.T., Miller, W.G. & Konkel, M.E. 2004. Differentiation of *Campylobacter coli*, *Campylobacter jejuni*, *Campylobacter* lari, and *Campylobacter* upsaliensis by a multiplex PCR developed from the nucleotide sequence of the lipid A gene lpxA. *Journal of Clinical Microbiology*, 42, 5549–5557, 10.1128/JCM.42.12.5549-5557.2004.

Konkel, M.E., Kim, B.J., Rivera-Amill, V. & Garvis, S.G. 1999. Bacterial secreted proteins are required for the internalization of *Campylobacter jejuni* into cultured mammalian cells. *Molecular Microbiology*, 32, 691–701, 10.1046/j.1365-2958.1999.01376.x.

Kopecko, D.J., Hu, L. & Zaal, K.J.M. 2001. *Campylobacter jejuni*-microtubule-dependent invasion. *Trends in Microbiology*, 9, 389–396, 10.1016/S0966-842X(01)02107-2.

Korlath, J.A., Osterholm, M.T., Judy, L.A., Forfang, J.C. & Robinson, R.A. 1985. A point-source outbreak of campylobacteriosis associated with consumption of raw milk. *Journal of Infectious Diseases*, 152, 592–596, 10.1093/infdis/152.3.592.

Korman, T.M., Varley, C.C. & Spelman, D.W. 1997. Acute hepatitis associated with *Campylobacter jejuni* bacteraemia. *Eur J Clin Microbiol Infect Dis*, 16, 678–681.

Korolik, V., Ketley, J. & Korolik, V. 2008. Chemosensory Signal Transduction Pathway of *Campylobacter jejuni*. *In Campylobacter, Third Edition*. American Society of Microbiology, 351–366., 10.1128/9781555815554.ch20.

Köser, C.U., Fraser, L.J., Ioannou, A., Becq, J., Ellington, M.J., Holden, M.T.G., Reuter, S. 2014. Rapid single-colony whole-genome sequencing of bacterial pathogens. *Journal of Antimicrobial Chemotherapy*, 69, 1275–1281, 10.1093/jac/dkt494.

Kothary, M.H. & Babu, U.S. 2001. Infective Dose of Foodborne Pathogens in Volunteers: a Review. *Journal of Food Safety*, 21, 49–68, 10.1111/j.1745-4565.2001.tb00307.x.

Kotloff, K.L., Nataro, J.P., Blackwelder, W.C., Nasrin, D., Farag, T.H.,

Panchalingam, S., Wu, Y. 2013. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): A prospective, case-control study. *The Lancet*, 382, 209–222, 10.1016/S0140-6736(13)60844-2.

Kovanen, S.M., Kivistö, R.I., Rossi, M., Schott, T., Kärkkäinen, U.M., Tuuminen, T., Uksila, J., Rautelin, H. & Hänninen, M.L. 2014. Multilocus sequence typing (MLST) and whole-genome MLST of *Campylobacter jejuni* isolates from human infections in three districts during a seasonal peak in Finland. *Journal of Clinical Microbiology*, 52, 4147–4154, 10.1128/JCM.01959-14.

Kusters, J.G., Van Vliet, A.H.M. & Kuipers, E.J. 2006. Pathogenesis of Helicobacter pylori infection. *Clinical Microbiology Reviews*, 19, 449–490, 10.1128/CMR.00054-05.

Langmead, B. & Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9, 357–359, 10.1038/nmeth.1923.

Lango-Scholey, L., Aidley, J., Woodacre, A., Jones, M.A. & Bayliss, C.D. 2016. High throughput method for analysis of repeat number for 28 phase variable loci of *Campylobacter jejuni* strain NCTC11168 Gerlach, R.G., *ed*. *PLoS ONE*, 11, e0159634, 10.1371/journal.pone.0159634.

Larsen, I.K., Gradel, K.O., Helms, M., Hornstrup, M.K., Jürgens, G., Mens, H., Rosager, C.L., Clausen, T.H., Kronborg, G. & Nielsen, H. 2011. Non-typhoidal Salmonella and *Campylobacter* infections among HIV-positive patients in Denmark. *Scandinavian Journal of Infectious Diseases*, 43, 3–7, 10.3109/00365548.2010.517780.

Lastovica, A.J. & Allos, B.M. 2008. Clinical Significance of *Campylobacter* and Related Species Other Than *Campylobacter jejuni* and *Campylobacter coli*. *In Campylobacter, 3rd Edition*. American Society of Microbiology, 123–149., 10.1128/9781555815554.ch7.

Lawson, G.H.K., Leaver, J.L., Pettigrew, G.W. & Rowland, A.C. 1981. Some Features of *Campylobacter* sputorum subsp. mucosalis subsp. nov.,

nom. rev. and Their Taxonomic Significance. *International journal of systematic bacteriology*, 31, 385–391, 10.1099/00207713-31-4-385.

Lázaro, B., Cárcamo, J., Audícana, A., Perales, I. & Fernández-Astorga, A. 1999. Viability and DNA maintenance in nonculturable spiral *Campylobacter jejuni* cells after long-term exposure to low temperatures. *Applied and Environmental Microbiology*, 65, 4677–4681.

Lee, G., Pan, W., Peñataro Yori, P., Paredes Olortegui, M., Tilley, D., Gregory, M., Oberhelman, R., Burga, R., Chavez, C.B. & Kosek, M. 2013. Symptomatic and Asymptomatic *Campylobacter* Infections Associated with Reduced Growth in Peruvian Children Vinetz, J.M., *ed*. *PLoS Neglected Tropical Diseases*, 7, e2036, 10.1371/journal.pntd.0002036.

LeJeune, J.T. & Rajala-Schultz, P.J. 2009. Unpasteurized Milk: A Continued Public Health Threat. *Clinical Infectious Diseases*, 48, 93–100, 10.1086/595007.

Lertsethtakarn, P., Ottemann, K.M. & Hendrixson, D.R. 2011. Motility and chemotaxis in *Campylobacter* and Helicobacter. *Annual Review of Microbiology*, 65, 389–410, 10.1146/annurev-micro-090110-102908.

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

Levin, R.E. 2007. *Campylobacter jejuni* : A Review of its Characteristics, Pathogenicity, Ecology, Distribution, Subspecies Characterization and Molecular Methods of Detection. *Food Biotechnology*, 21, 271–347, 10.1080/08905430701536565.

Levy, A.J. 1946. A gastro-enteritis cutbreak probably due to a bovine strain of vibrio. *The Yale journal of biology and medicine*, 18, 243–258.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078–2079, 10.1093/bioinformatics/btp352.

Li, Z., Lou, H., Ojcius, D.M., Sun, A., Sun, D., Zhao, J., Lin, X. & Yan, J. 2014. Methyl-accepting chemotaxis proteins 3 and 4 are responsible for *Campylobacter jejuni* chemotaxis and jejuna colonization in mice in response to sodium deoxycholate. *Journal of Medical Microbiology*, 63, 343–354, 10.1099/jmm.0.068023-0.

Lieberman, T.D., Flett, K.B., Yelin, I., Martin, T.R., McAdam, A.J., Priebe, G.P. & Kishony, R. 2013. Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nature Genetics*, 46, 82–87, 10.1038/ng.2848.

Lindén, S.K., Florin, T.H.J. & McGuckin, M.A. 2008. Mucin dynamics in intestinal bacterial infection Gay, N., *ed*. *PLoS ONE*, 3, e3952, 10.1371/journal.pone.0003952.

Linton, D., Owen, R.J. & Stanley, J. 1996. Rapid identification by PCR of the genus *Campylobacter* and of five *Campylobacter* species enteropathogenic for man and animals. *Research in Microbiology*, 147, 707–718, 10.1016/S0923-2508(97)85118-2.

Linton, D., Lawson, A., Owen, R. & Stanley, J. 1997. PCR detection , identification to species level , and fingerprinting of *Campylobacter jejuni* and *Campylobacter coli* direct from diarrheic samples. *Journal of Clinical*, 35, 2568–2572.

Louis, V.R., Gillespie, I.A., O'Brien, S.J., Russek-Cohen, E., Pearson, A.D. & Colwell, R.R. 2005. Temperature-driven *Campylobacter* seasonality in England and Wales. *Applied and Environmental Microbiology*, 71, 85–92, 10.1128/AEM.71.1.85-92.2005.

Lous, J., Ryborg, C.T. & Thomsen, J.L. 2011. Bergey Manual. *International Journal of Pediatric Otorhinolaryngology*, 75, 1058–1061, 10.1016/j.ijporl.2011.05.009.

Lugert, R., Gross, U. & Zautner, A.E. 2015. *Campylobacter jejuni*: components for adherence to and invasion of eukaryotic cells. *Berliner und Münchener Tierärztliche Wochenschrift*, 128, 90–97, 10.2376/0005-9366-128-10.

MacFaddin, J.F. 1980. Biochemical Tests for Identification of Medical

Bacteria. *Biochemical Tests for Identification of Medical Bacteria. 2nd edition*, 527, 10.1136/jcp.29.10.958-c.

Maiden, M.C.J., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q. 1998. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences*, 95, 3140–3145, 10.1073/pnas.95.6.3140.

Marchant, J., Wren, B. & Ketley, J. 2002. Exploiting genome sequence: Predictions for mechanisms of *Campylobacter* chemotaxis. *Trends in Microbiology*, 10, 155–159, 10.1016/S0966-842X(02)02323-5.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J. 2006. Corrigendum: Genome sequencing in microfabricated high-density pi*coli*tre reactors. *Nature*, 441, 120–120, 10.1038/nature04726.

McGrogan, A., Madle, G.C., Seaman, H.E. & De Vries, C.S. 2009. The epidemiology of Guillain-Barré syndrome worldwide: A systematic literature review. *Neuroepidemiology*, 32, 150–163, 10.1159/000184748.

Meldrum, R.J., Griffiths, J.K., Smith, R.M.M. & Evans, M.R. 2005. The seasonality of human *Campylobacter* infection and *Campylobacter* isolates from fresh, retail chicken in Wales. *Epidemiology and infection*, 133, 49–52, 10.1017/S0950268804003188.

Miller, C.E., Williams, P.H. & Ketley, J.M. 2009. Pumping iron: Mechanisms for iron uptake by *Campylobacter*. *Microbiology*, 155, 3157–3165, 10.1099/mic.0.032425-0.

Mizrahi-Man, O., Davenport, E.R. & Gilad, Y. 2013. Taxonomic Classification of Bacterial 16S rRNA Genes Using Short Sequencing Reads: Evaluation of Effective Study Designs. *PLoS ONE*, 8, e53608, 10.1371/journal.pone.0053608.

Mohawk, K.L., Poly, F., Sahl, J.W., Rasko, D.A. & Guerry, P. 2014. High frequency, spontaneous motA mutations in *Campylobacter jejuni* strain 81-176. *PLoS ONE*, 9, e88043, 10.1371/journal.pone.0088043.

Moore, J.E. 2001. Bacterial dormancy in *Campylobacter*: Abstract theory or cause for concern? *International Journal of Food Science and Technology*, 36, 593–600, 10.1046/j.1365-2621.2001.00508.x.

Mossong, J., Mughini-Gras, L., Penny, C., Devaux, A., Olinger, C., Losch, S., Cauchie, H.-M., van Pelt, W. & Ragimbeau, C. 2016. Human campylobacteriosis in Luxembourg, 2010-2013: A Case-Control Study Combined with Multilocus Sequence Typing for Source Attribution and Risk Factor Analysis. *Scientific reports*, 6, 20939, 10.1038/srep20939.

Murphy, C., Carroll, C. & Jordan, K.N. 2006. Environmental survival mechanisms of the foodborne pathogen *Campylobacter jejuni*. *Journal of Applied Microbiology*, 100, 623–632, 10.1111/j.1365-2672.2006.02903.x.

Mwangi, M.M., Wu, S.W., Zhou, Y., Sieradzki, K., de Lencastre, H., Richardson, P., Bruce, D. 2007. Tracking the in vivo evolution of multidrug resistance in Staphylococcus aureus by whole-genome sequencing. *Proceedings of the National Academy of Sciences*, 104, 9451–9456, 10.1073/pnas.0609839104.

Nachamkin, I., Allos, B.M. & Ho, T. 1998. *Campylobacter* species and Guillain-Barré syndrome. *Clinical microbiology reviews*, 11, 555–567.

Naikare, H., Butcher, J., Flint, A., Xu, J., Raymond, K.N. & Stintzi, A. 2013. *Campylobacter jejuni* ferric–enterobactin receptor CfrA is TonB3 dependent and mediates iron acquisition from structurally different catechol siderophores. *Metallomics*, 5, 988, 10.1039/c3mt20254b.

Neal, K.R., Scott, H.M., Slack, R.C. & Logan, R.F. 1996. Omeprazole as a risk factor for *Campylobacter* gastroenteritis: case-control study. *BMJ (Clinical research ed.)*, 312, 414–415, 10.1016/S0736-4679(97)86454-2.

Newell, D.G. & Fearnley, C. 2003. Sources of *Campylobacter* colonization in broiler chickens. *Applied and Environmental Microbiology*, 69, 4343–4351, 10.1128/AEM.69.8.4343-4351.2003.

Newell, D.G., Elvers, K.T., Dopfer, D., Hansson, I., Jones, P., James, S., Gittins, J. 2011. Biosecurity-based interventions and strategies to reduce *Campylobacter* spp. on poultry farms. *Applied and Environmental Microbiology*, 77, 8605–8614, 10.1128/AEM.01090-10.

Nguyen, D. & Singh, P.K. 2006. Evolving stealth: genetic adaptation of Pseudomonas aeruginosa during cystic fibrosis infections. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 8305–8306, 10.1073/pnas.0602526103.

Nichols, G.L., Richardson, J.F., Sheppard, S.K., Lane, C. & Sarran, C. 2012. *Campylobacter* epidemiology: a descriptive study reviewing 1 million cases in England and Wales between 1989 and 2011. *BMJ Open*, 2, e001179, 10.1136/bmjopen-2012-001179.

Nielsen, L.N., Luijkx, T.A., Vegge, C.S., Johnsen, C.K., Nuijten, P., Wren, B.W., Ingmer, H. & Krogfelt, K.A. 2012. Identification of immunogenic and virulence-associated *Campylobacter jejuni* proteins. *Clinical and Vaccine Immunology*, 19, 113–119, 10.1128/CVI.05161-11.

Noormohamed, A. & Fakhr, M.K. 2013. A higher prevalence rate of *Campylobacter* in retail beef livers compared to other beef and pork meat cuts. *International Journal of Environmental Research and Public Health*, 10, 2058–2068, 10.3390/ijerph10052058.

Nuijten, P.J.M., Van der Zeijst, B.A.M. & Newell, D.G. 1991. Localization of immunogenic regions on the flagellin proteins of *Campylobacter jejuni* 81116. *Infection and Immunity*, 59, 1100–1105.

O Cróinín, T. & Backert, S. 2012. Host epithelial cell invasion by *Campylobacter jejuni*: trigger or zipper mechanism? *Frontiers in cellular and infection microbiology*, 2, 25, 10.3389/fcimb.2012.00025.

Obiri-Danso, K. & Jones, K. 1999. The effect of a new sewage treatment plant on faecal indicator numbers, *Campylobacter*s and bathing water compliance in Morecambe Bay. *Journal of Applied Microbiology*, 86, 603–614, 10.1046/j.1365-2672.1999.00703.x.

Oliver, A., Cantón, R., Campo, P., Baquero, F., Blázquez, J., Tsui, L.C.,

Rosenstein, B.J. 2000. High frequency of hypermutable Pseudomonas aeruginosa in cystic fibrosis lung infection. *Science (New York, N.Y.)*, 288, 1251–1254, 10.1126/science.288.5469.1251.

Olson, C.K., Ethelberg, S., van Pelt, W. & Tauxe, R. V. 2008. Epidemiology of *Campylobacter jejuni* Infections in Industrialized Nations. *In Campylobacter, Third Edition*. American Society of Microbiology, 163–189., 10.1128/9781555815554.ch9.

Olson, N.D., Lund, S.P., Colman, R.E., Foster, J.T., Sahl, J.W., Schupp, J.M., Keim, P., Morrow, J.B., Salit, M.L. & Zook, J.M. 2015. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Frontiers in Genetics*, 6, 235, 10.3389/fgene.2015.00235.

On, S.L. 2001. Taxonomy of *Campylobacter*, Arcobacter, Helicobacter and related bacteria: current status, future prospects and immediate concerns. *Symposium series (Society for Applied Microbiology)*, 90, 1S–15S, 10.1046/j.1365-2672.2001.01349.x.

On, S.L.W. 2013. Isolation, identification and subtyping of *Campylobacter*: Where to from here? *Journal of Microbiological Methods*, 95, 3–7, 10.1016/j.mimet.2013.06.011.

Op den Winkel, M., Gulberg, V., Weiss, M., Ebeling, F., Gerbes, A.L. & Samtleben, W. 2010. Acute postinfectious glomerulonephritis associated with *Campylobacter jejuni* enteritis - a case report and review of the literature on C. *jejuni*'s potential to trigger immunologically mediated renal disease. *Clin Nephrol*, 74, 474–479, 8149 [pii].

Osano, O. & Arimi, S.M. 1999. Retail poultry and beef as sources of *Campylobacter jejuni*. *East African Medical Journal*, 76, 141–143.

Pacanowski, J., Lalande, V., Lacombe, K., Boudraa, C., Lesprit, P., Legrand, P., Trystram, D. 2008. *Campylobacter* Bacteremia: Clinical Features and Factors Associated with Fatal Outcome. *Clinical Infectious Diseases*, 47, 790–796, 10.1086/591530.

Padungton, P. & Kaneene, J.B. 2003. *Campylobacter* spp in human,

chickens, pigs and their antimicrobial resistance. *The Journal of veterinary medical science / the Japanese Society of Veterinary Science*, 65, 161–170, 10.1292/jvms.65.161.

Pallen, M.J. & Wren, B.W. 2007. Bacterial pathogenomics. *Nature*, 449, 835–842, 10.1038/nature06248.

Parkhill, J., Achtman, M., James, K.D., Bentley, S.D., Churcher, C., Klee, S.R., Morelli, G. 2000a. Complete DNA sequence of a serogroup A strain of Neisseria meningitidis Z2491. *Nature*, 404, 502–506, 10.1038/35006655.

Parkhill, J., Wren, B.W., Mungall, K., Ketley, J.M., Churcher, C., Basham, D., Chillingworth, T. 2000b. The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*, 403, 665–668, 10.1038/35001088.

Pearson, A D., Greenwood, M.H., Donaldson, J., Healing, T.D., Jones, D.M., Shahamat, M., Feltham, R.K. & Colwell, R.R. 2000. Continuous source outbreak of campylobacteriosis traced to chicken. *Journal of food protection*, 63, 309–314.

Pearson, B.M., Gaskin, D.J.H., Segers, R.P.A.M., Wells, J.M., Nuijten, P.J.M. & Van Vliet, A.H.M. 2007. The complete genome sequence of *Campylobacter jejuni* strain 81116 (NCTC11828). *Journal of Bacteriology*, 189, 8402–8403, 10.1128/JB.01404-07.

PHE & DEFRA. 2013. Zoonoses Report 2012. 31–32.

Phillips, I. 1993. Cowan and Steel's Manual for the Identification of Medical Bacteria. *Journal of Clinical Pathology*, 46, 975–975, 10.1136/jcp.46.10.975-a.

Phongsisay, V., Perera, V.N. & Fry, B.N. 2007. Expression of the htrB gene is essential for responsiveness of Salmonella typhimurium and *Campylobacter jejuni* to harsh environments. *Microbiology*, 153, 254–262, 10.1099/mic.0.29230-0.

Piccirillo, A., Niero, G., Calleros, L., Pérez, R., Naya, H. & Iraola, G. 2016. *Campylobacter* geochelonis sp. nov. isolated from the western Hermann's tortoise (Testudo hermanni hermanni). *International*

Journal of Systematic and Evolutionary Microbiology, 66, 3468–3476, 10.1099/ijsem.0.001219.

Poly, F., Read, T., Tribble, D.R., Baqar, S., Lorenzo, M. & Guerry, P. 2007b. Genome sequence of a clinical isolate of *Campylobacter jejuni* from Thailand. *Infection and Immunity*, 75, 3425–3433, 10.1128/IAI.00050-07.

Poly, F. & Guerry, P. 1989. Pathogenesis of *Campylobacter*. 66, 629–630.

Poly, F., Ewing, C., Goon, S., Hickey, T.E., Rockabrand, D., Majam, G., Lee, L., Phan, J., Savarino, N.J. & Guerry, P. 2007. Heterogeneity of a *Campylobacter jejuni* protein that is secreted through the flagellar filament. *Infection and Immunity*, 75, 3859–3867, 10.1128/IAI.00159-07.

Porter, I.A. & Reid, T.M. 1980. A milk-borne outbreak of *Campylobacter* infection. *Journal of Hygiene (London)*, 84, 415–419, 10.1017/S0022172400026942.

Power, M.E., Guerry, P., McCubbin, W.D., Kay, C.M. & Trust, T.J. 1994. Structural and antigenic characteristics of *Campylobacter coli* FlaA flagellin. *Journal of Bacteriology*, 176, 3303–3313.

Rahman, H., King, R.M., Shewell, L.K., Semchenko, E.A., Hartley-Tassell, L.E., Wilson, J.C., Day, C.J. & Korolik, V. 2014. Characterisation of a multi-ligand binding chemoreceptor CcmL (Tlp3) of *Campylobacter jejuni*. *PLoS pathogens*, 10, e1003822, 10.1371/journal.ppat.1003822.

Ramirez-Hernandez, A., Rupnow, J. & Hutkins, R.W. 2015. Adherence Reduction of *Campylobacter jejuni* and *Campylobacter coli* Strains to HEp-2 Cells by Mannan Oligosaccharides and a High-Molecular-Weight Component of Cranberry Extract. *Journal of food protection*, 78, 1496–1505, 10.4315/0362-028X.JFP-15-087.

Rasko, D.A., Worsham, P.L., Abshire, T.G., Stanley, S.T., Bannan, J.D., Wilson, M.R., Langham, R.J. 2011. Bacillus anthracis comparative genome analysis in support of the Amerithrax investigation. *Proceedings of the National Academy of Sciences*, 108, 5027–5032, 10.1073/pnas.1016657108.

Read, T.D., Salzberg, S.L., Pop, M., Shumway, M., Umayam, L., Jiang, L., Holtzapple, E. 2002. Comparative genome sequencing for discovery of novel polymorphisms in Bacillus anthracis. *Science (New York, N.Y.)*, 296, 2028–2033, 10.1126/science.1071837.

Reuter, M. & van Vliet, A.H.M. 2013. Signal Balancing by the CetABC and CetZ Chemoreceptors Controls Energy Taxis in *Campylobacter jejuni*. *PLoS ONE*, 8, 1–10, 10.1371/journal.pone.0054390.

Revez, J., Schott, T., Llarena, A.K., Rossi, M. & Hänninen, M.L. 2013. Genetic heterogeneity of *Campylobacter jejuni* NCTC 11168 upon human infection. *Infection, Genetics and Evolution*, 16, 305–309, 10.1016/j.meegid.2013.03.009.

Revez, J., Zhang, J., Schott, T., Kivistö, R., Rossi, M. & Hänninen, M.L. 2014. Genomic variation between *Campylobacter jejuni* isolates associated with milk-borne-disease outbreaks. *Journal of Clinical Microbiology*, 52, 2782–2786, 10.1128/JCM.00931-14.

Rosner, B.M., Stark, K. & Werber, D. 2010. Epidemiology of reported Yersinia entero*coli*tica infections in Germany, 2001-2008. *BMC public health*, 10, 337, 10.1186/1471-2458-10-337.

Rotariu, O., Dallas, J.F., Ogden, I.D., MacRae, M., Sheppard, S.K., Maiden, M.C.J., Gormley, F.J., Forbes, K.J. & Strachan, N.J.C. 2009. Spatiotemporal Homogeneity of *Campylobacter* Subtypes from Cattle and Sheep across Northeastern and Southwestern Scotland. *Applied and Environmental Microbiology*, 75, 6275–6281, 10.1128/AEM.00499-09.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.-A. & Barrell, B. 2000. Artemis: sequence visualization and annotation. *Bioinformatics*, 16, 944–945, 10.1093/bioinformatics/16.10.944.

Saha, S.K., Saha, S. & Sanyal, S.C. 1991. Recovery of injured *Campylobacter jejuni* cells after animal passage. *Applied and Environmental Microbiology*, 57, 3388–3389.

Scallan, E., Hoekstra, R.M., Angulo, F.J., Tauxe, R. V, Widdowson, M.A.,

Roy, S.L., Jones, J.L. & Griffin, P.M. 2011. Foodborne illness acquired in the United States-Major pathogens. *Emerging Infectious Diseases*, 17, 7–15, 10.3201/eid1701.P11101.

Schielke, A., Rosner, B.M. & Stark, K. 2014. Epidemiology of campylobacteriosis in Germany - insights from 10 years of surveillance. *BMC infectious diseases*, 14, 30, 10.1186/1471-2334-14-30.

Schönberg-Norio, D., Takkinen, J., Hänninen, M.L., Katila, M.L., Kaukoranta, S.S., Mattila, L. & Rautelin, H. 2004. Swimming and *Campylobacter* infections. *Emerging Infectious Diseases*, 10, 1474–1477, 10.3201/eid1008.030924.

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

Senes, A., Gerstein, M. & Engelman, D.M. 2000. Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with β-branched residues at neighboring positions. *Journal of Molecular Biology*, 296, 921–936, 10.1006/jmbi.1999.3488.

Shahrizaila, N. & Yuki, N. 2016. Guillain-barré syndrome and *Campylobacter jejuni* enteritis. *In Neuro-Immuno-Gastroenterology*. Cham: Springer International Publishing, 169–180., 10.1007/978-3-319-28609-9_9.

Shandera, W.X., Tormey, M.P. & Blaser, M.J. 1992. An outbreak of bacteremic *Campylobacter jejuni* infection. *Mt Sinai Journal of Medicine*, 59, 53–56.

Sheppard, S.K., McCarthy, N.D., Falush, D. & Maiden, M.C.J. 2008. Convergence of *Campylobacter* species: implications for bacterial evolution. *Science (New York, N.Y.)*, 320, 237–239, 10.1126/science.1155532.

Sheppard, S.K., Dallas, J.F., MacRae, M., McCarthy, N.D., Sproston, E.L., Gormley, F.J., Strachan, N.J.C., Ogden, I.D., Maiden, M.C.J. & Forbes, K.J. 2009. *Campylobacter* genotypes from food animals, environmental sources and clinical disease in Scotland 2005/6.

International Journal of Food Microbiology, 134, 96–103, 10.1016/j.ijfoodmicro.2009.02.010.

Sheppard, S.K., Cheng, L., Méric, G., De Haan, C.P.A., Llarena, A.K., Marttinen, P., Vidal, A. 2014. Cryptic ecology among host generalist *Campylobacter jejuni* in domestic animals. *Molecular Ecology*, 23, 2442–2451, 10.1111/mec.12742.

Sheppard, S.K. & Maiden, M.C.J. 2015. The evolution of *Campylobacter jejuni* and *Campylobacter coli*. *Cold Spring Harbor Perspectives in Biology*, 7, a018119, 10.1101/cshperspect.a018119.

Shoaf-Sweeney, K.D., Larson, C.L., Tang, X. & Konkel, M.E. 2008. Identification of *Campylobacter jejuni* proteins recognized by maternal antibodies of chickens. *Applied and Environmental Microbiology*, 74, 6867–6875, 10.1128/AEM.01097-08.

Silva, J., Leite, D., Fernandes, M., Mena, C., Gibbs, P.A. & Teixeira, P. 2011. *Campylobacter* spp. As a foodborne pathogen: A review. *Frontiers in Microbiology*, 2, 200, 10.3389/fmicb.2011.00200.

Skarp, C.P.A., Hönninen, M.L. & Rautelin, H.I.K. 2016. Campylobacteriosis: The role of poultry meat. *Clinical Microbiology and Infection*, 22, 103–109, 10.1016/j.cmi.2015.11.019.

Skirrow, M.B. 1977. *Campylobacter* enteritis: a "new" disease. *British medical journal*, 2, 9–11.

Skirrow, M.B. & Benjamin, J. 1980. "1001" *Campylobacter*s: cultural characteristics of intestinal *Campylobacter*s from man and animals. *The Journal of hygiene*, 85, 427–442, 10.1017/S0022172400063506.

Skirrow, M.B., Jones, D.M., Sutcliffe, E. & Benjamin, J. 1993. *Campylobacter* bacteraemia in England and Wales, 1981-91. *Epidemiology and infection*, 110, 567–573, 10.1017/S0950268800050986.

Skirrow, M.B. 2006. John McFadyean and the centenary of the first isolation of *Campylobacter* species. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 43, 1213–1217, 10.1086/508201.

Smith, E.E., Buckley, D.G., Wu, Z., Saenphimmachak, C., Hoffman, L.R., D'Argenio, D.A., Miller, S.I. 2006. Genetic adaptation by Pseudomonas aeruginosa to the airways of cystic fibrosis patients. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 8487–8492, 10.1073/pnas.0602138103.

Smith, K.D., Andersen-Nissen, E., Hayashi, F., Strobe, K., Bergman, M.A., Barrett, S.L.R., Cookson, B.T. & Aderem, A. 2003. Toll-like receptor 5 recognizes a conserved site on flagellin required for protofilament formation and bacterial motility. *Nature Immunology*, 4, 1247–1253, 10.1038/ni1011.

Smith, T. & Taylor, M.S. 1919. Some Morphological and Biological Characters of the Spirilla (Vibrio Fetus, N. Sp.) Associated With Disease of the Fetal Membranes in Cattle. *The Journal of experimental medicine*, 30, 299–311, 10.1084/jem.30.4.299.

Snitkin, E.S., Zelazny, A.M., Thomas, P.J., Stock, F., Henderson, D.K., Palmore, T.N. & Segre, J.A. 2012. Tracking a Hospital Outbreak of Carbapenem-Resistant Klebsiella pneumoniae with Whole-Genome Sequencing. *Science Translational Medicine*, 4, 148ra116-148ra116, 10.1126/scitranslmed.3004129.

Sorvillo, F.J., Lieb, L.E. & Waterman, S.H. 1991. Incidence of campylobacteriosis among patients with AIDS in Los Angeles County. *Journal of Acquiedr Immune Deficiency Syndrome*, 4, 598–602.

Sparling, P.F. 1966. Genetic transformation of Neisseria gonorrhoeae to streptomycin resistance. *Journal of Bacteriology*, 92, 1364–1371.

Stahl, M., Butcher, J. & Stintzi, A. 2012. Nutrient acquisition and metabolism by *Campylobacter jejuni*. *Frontiers in cellular and infection microbiology*, 2, 5, 10.3389/fcimb.2012.00005.

Stanley, K. & Jones, K. 2003. Cattle and sheep farms as reservoirs of *Campylobacter*. *Journal of applied microbiology*, 94, 104S–113S, 10.1046/j.1365-2672.94.s1.12.x.

Stathopoulos, C. 1999. Bacterial outer membrane proteins: topological

analyses and biotechnological perspectives. *Membrane Cell Biology*, 13, 3–21.

Stern, N.J., Fedorka-Cray, P., Bailey, J.S., Cox, N.A., Craven, S.E., Hiett, K.L., Musgrove, M.T., Ladely, S., Cosby, D. & Mead, G.C. 2001. Distribution of *Campylobacter* spp. in selected U.S. poultry production and processing operations. *Journal of food protection*, 64, 1705–1710.

Stern, N.J., Hiett, K.L., Alfredsson, G.A., Kristinsson, K.G., Reiersen, J., Hardardottir, H., Briem, H. 2003. *Campylobacter* spp. in Icelandic poultry operations and human disease. *Epidemiology and infection*, 130, 23–32, 10.1017/S0950268802007914.

Strachan, N.J.C., Gormley, F.J., Rotariu, O., Ogden, I.D., Miller, G., Dunn, G.M., Sheppard, S.K. 2009. Attribution of *Campylobacter* Infections in Northeast Scotland to Specific Sources by Use of Multilocus Sequence Typing. *The Journal of Infectious Diseases*, 199, 1205–1208, 10.1086/597417.

Studahl, A. & Andersson, Y. 2000. Risk factors for indigenous *Campylobacter* infection: a Swedish case-control study. *Epidemiology and infection*, 125, 269–275, 10.1017/S0950268899004562.

Szymanski, C.M., King, M., Haardt, M. & Armstrong, G.D. 1995. *Campylobacter jejuni* motility and invasion of Caco-2 cells. *Infection and Immunity*, 63, 4295–4300.

Szymanski, C.M., Logan, S.M., Linton, D. & Wren, B.W. 2003. *Campylobacter* - A tale of two protein glycosylation systems. *Trends in Microbiology*, 11, 233–238, 10.1016/S0966-842X(03)00079-9.

Tam, C.C., Rodrigues, L.C., O'Brien, S.J. & Hajat, S. 2006. Temperature dependence of reported *Campylobacter* infection in England, 1989-1999. *Epidemiol Infect*, 134, 119–125, 10.1017/S0950268805004899.

Tam, C.C., Rodrigues, L.C., Viviani, L., Dodds, J.P., Evans, M.R., Hunter, P.R., Gray, J.J. 2012. Longitudinal study of infectious intestinal disease in the UK (IID2 study): incidence in the community and

presenting to general practice. *Gut*, 61, 69–77, 10.1136/gut.2011.238386.

Tanner, A.C.R., Badger, S., Lai, C.-H., Listgarten, M.A., Visconti, R.A. & Socransky ', S.S. 1981. Eikenella corrodens from Humans with Periodontal Disease. *International Journal of Systematic Bacteriology*, 432–445.

Taylor, E. V., Herman, K.M., Ailes, E.C., Fitzgerald, C., Yoder, J.S., Mahon, B.E. & Tauxe, R. V. 2012. Common source outbreaks of *Campylobacter* infection in the USA, 1997–2008. *Epidemiology and infection*, 141, 1–10, 10.1017/S0950268812001744.

Tee, W. & Mijch, a. 1998. *Campylobacter jejuni* bacteremia in human immunodeficiency virus (HIV)-infected and non-HIV-infected patients: comparison of clinical features and review. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 26, 91–96, Doi 10.1086/516263.

Thibault, P., Logan, S.M., Kelly, J.F., Brisson, J.R., Ewing, C.P., Trust, T.J. & Guerry, P. 2001. Identification of the Carbohydrate Moieties and Glycosylation Motifs in *Campylobacter jejuni* Flagellin. *Journal of Biological Chemistry*, 276, 34862–34870, 10.1074/jbc.M104529200.

Thomas, D.K., Lone, A.G., Brent Selinger, L., Taboada, E.N., Uwiera, R.R.E., Wade Abbott, D. & Douglas Inglis, G. 2014. Comparative variation within the genome of *Campylobacter jejuni* NCTC 11168 in human and murine hosts Freitag, N.E., *ed*. *PLoS ONE*, 9, e88229, 10.1371/journal.pone.0088229.

Totten, P.A., Fennell, C.L., Tenover, F.C., Wezenberg, J.M., Perine, P.L. & Stamm, W.E. 1985. *Campylobacter* cinaedi (sp. nov.) and *Campylobacter* fennelliae (sp, nov.): Two new *Campylobacter* species associated with enteric disease in homosexual men. *Journal of Infectious Diseases*, 151, 42–46, 10.1093/infdis/151.1.131.

Tracz, D.M., Keelan, M., Ahmed-Bentley, J., Gibreel, A., Kowalewska-Grochowska, K. & Taylor, D.E. 2005. pVir and bloody diarrhea in *Campylobacter jejuni* enteritis. *Emerging Infectious Diseases*, 11, 838–

843, 10.3201/eid1106.041052.

Treangen, T.J., Ondov, B.D., Koren, S. & Phillippy, A.M. 2014. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology*, 15, 524, 10.1186/s13059-014-0524-x.

Tsuda, M., Karita, M. & Nakazawa, T. 1993. Genetic transformation in Helicobacter pylori. *Microbiology and immunology*, 37, 85–89.

Udayakumar, D. & Sanaullah, M. 2009. *Campylobacter* cholecystitis. *International Journal of Medical Sciences*, 6, 374–375, 10.1016/S0140-6736(79)92997-0.

Urumova, V., Stoyanchev, T., Lyutskanov, M., Daskalov, H., Vashin, I. & Maramski, A. 2014. Antimicrobial sensitivity of *Campylobacter jejuni* poultry isolates from the Republic of Bulgaria. *Istanbul Universitesi Veteriner Fakultesi Dergisi*, 40, 29–34.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K. 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Research*, 18, 1051–1063, 10.1101/gr.076463.108.

Van, T.T.H., Elshagmani, E., Gor, M.C., Scott, P.C. & Moore, R.J. 2016. *Campylobacter* hepaticus sp. nov., isolated from chickens with spotty liver disease. *International Journal of Systematic and Evolutionary Microbiology*, 66, 4518–4524, 10.1099/ijsem.0.001383.

Van Dijk, E.L., Auger, H., Jaszczyszyn, Y. & Thermes, C. 2014. Ten years of next-generation sequencing technology. *Trends in Genetics*, 10.1016/j.tig.2014.07.001.

Vandamme, P., Daneshvar, M.I., Dewhirst, F.E., Paster, B.J., Kersters, K., Goossens, H. & Moss, C.W. 1995. Chemotaxonomic analyses of Bacteroides gracilis and Bacteroides ureolyticus and reclassification of B. gracilis as *Campylobacter* gracilis comb. nov. *International journal of systematic bacteriology*, 45, 145–152, 10.1099/00207713-45-1-145.

Vandamme, P., Falsen, E., Rossau, R., Hoste, B., Segers, P., Tytgat, R. &

De Ley, J. 1991. Revision of *Campylobacter*, Helicobacter, and Wolinella taxonomy: emendation of generic descriptions and proposal of Arcobacter gen. nov. *International journal of systematic bacteriology*, 41, 88–103, 10.1099/00207713-41-1-88.

Vandamme, P., Gevers, D. & Debruyne, L. 2008. Taxonomy of the Family *Campylobacter*aceae. *In Campylobacter, Third Edition*. American Society of Microbiology, 3–25., 10.1128/9781555815554.ch1.

Vellinga, A. & Van Loock, F. 2002. The dioxin crisis as experiment to determine poultry-related *Campylobacter* enteritis. *Emerging Infectious Diseases*, 8, 19–22, 10.3201/eid0801.010129.

Veron, M. & Chatelain, R. 1973. Taxonomic Study of the Genus *Campylobacter* Sebald and Veron and Designation of the Neotype Strain for the Type Species, *Campylobacter* fetus (Smith and Taylor) Sebald and Veron. *International Journal of Systematic Bacteriology*, 23, 122–134, 10.1099/00207713-23-2-122.

Voelkerding, K. V., Dames, S.A. & Durtschi, J.D. 2009. Next-generation sequencing:from basic research to diagnostics. *Clinical Chemistry*, 55, 641–658, 10.1373/clinchem.2008.112789.

Vogt, R.L., Sours, H.E., Barrett, T., Feldman, R.A., Dickinson, R.J. & Witherell, L. 1982. *Campylobacter* enteritis associated with contaminated water. *Annals of Internal Medicine*, 96, 292–296, 10.7326/0003-4819-96-3-292.

Wadhams, G.H. & Armitage, J.P. 2004. Making sense of it all: bacterial chemotaxis. *Nat.Rev.Mol.Cell Biol.*, 5, 1024–1037, 10.1038/nrm1524.

Waldenström, J., Axelsson-Olsson, D., Olsen, B., Hasselquist, D., Griekspoor, P., Jansson, L., Teneberg, S., Svensson, L. & Ellström, P. 2010. *Campylobacter jejuni* colonization in wild birds: Results from an infection experiment Ahmed, N., *ed*. *PLoS ONE*, 5, e9082, 10.1371/journal.pone.0009082.

Waldenström, J., Broman, T., Carlsson, I., Hasselquist, D., Achterberg, R.P., Wagenaar, J.A. & Olsen, B. 2002. Prevalence of *Campylobacter jejuni*, *Campylobacter* lari, and *Campylobacter coli* in different

ecological guilds and taxa of migrating birds. *Applied and Environmental Microbiology*, 68, 5911–5917, 10.1128/AEM.68.12.5911-5917.2002.

Walker, R.I., Caldwell, M.B., Lee, E.C., Guerry, P., Trust, T.J. & Ruiz-Palacios, G.M. 1986. Pathophysiology of *Campylobacter* enteritis. *Microbiological reviews*, 50, 81–94.

Wang, Q., Li, B., Tsang, A.K.L., Yi, Y., Woo, P.C.Y. & Liu, C.H. 2013. Genotypic Analysis of Klebsiella pneumoniae Isolates in a Beijing Hospital Reveals High Genetic Diversity and Clonal Population Structure of Drug-Resistant Isolates Manganelli, R., *ed*. *PLoS ONE*, 8, e57091, 10.1371/journal.pone.0057091.

Wassenaar, T.M. 1997. Toxin production by *Campylobacter* spp. *Clinical Microbiology Reviews*, 10, 466–476.

Wassenaar, T.M. & Blaser, M.J. 1999. Pathophysiology of *Campylobacter jejuni* infections of humans. *Microbes and Infection*, 1, 1023–1033, 10.1016/S1286-4579(99)80520-6.

Wassenaar, T.M., Bleumink-Pluym, N.M. & van der Zeijst, B.A. 1991. Inactivation of *Campylobacter jejuni* flagellin genes by homologous recombination demonstrates that flaA but not flaB is required for invasion. *The EMBO journal*, 10.

Wassenaar, T.M., Bleumink-Pluym, N.M.C., Newell, D.G., Nuijten, P.J.M. & Van der Zeijst, B.A.M. 1994. Differential flagellin expression in a flaA flaB+ mutant of *Campylobacter jejuni*. *Infection and Immunity*, 62, 3901–3906.

Wassenaar, T.M., Fry, B.N. & Vanderzeijst, B.A.M. 1995. Variation of the Flagellin Gene Locus of *Campylobacter-Jejuni* by Recombination and Horizontal Gene-Transfer. *Microbiology-Uk*, 141, 95–101.

Wassenaar, T.M., Van Der Zeijst, B.A., Ayling, R. & Newell, D.G. 1993. Colonization of chicks by motility mutants of *Campylobacter jejuni* demonstrates the importance of flagellin A expression. *Journal of General Microbiology*, 139, 1171–1175, 10.1099/00221287-139-6-1171.

Wells, J.M. & Bennik, M.H.J. 2003. Genomics of food-borne bacterial

pathogens. 21–35, 10.1079/NRR200358.

Wesley, I. V., Wells, S.J., Harmon, K.M., Green, A., Schroeder-Tucker, L., Glover, M. & Siddique, I. 2000. Fecal shedding of *Campylobacter* and Arcobacter spp. in dairy cattle. *Applied and Environmental Microbiology*, 66, 1994–2000, 10.1128/AEM.66.5.1994-2000.2000.

Wilson, D.J., Gabriel, E., Leatherbarrow, A.J.H., Cheesbrough, J., Gee, S., Bolton, E., Fox, A., Fearnhead, P., Hart, C.A. & Diggle, P.J. 2008. Tracing the source of campylobacteriosis Guttman, D.S., *ed. PLoS Genetics*, 4, e1000203, 10.1371/journal.pgen.1000203.

Wingstrand, A., Neimann, J., Engberg, J., Nielsen, E.M., Gerner-Smidt, P., Wegener, H.C. & Mølbak, K. 2006. Fresh chicken as main risk factor for campylobacteriosis, Denmark. *Emerging Infectious Diseases*, 12, 280–284, 10.3201/eid1202.050936.

Wood, R.C. 1992. *Campylobacter* Enteritis Outbreaks Associated With Drinking Raw Milk During Youth Activities. *JAMA*, 268, 3228, 10.1001/jama.1992.03490220072031.

Workentine, M.L., Sibley, C.D., Glezerson, B., Purighalla, S., Norgaard-Gron, J.C., Parkins, M.D., Rabin, H.R. & Surette, M.G. 2013. Phenotypic Heterogeneity of Pseudomonas aeruginosa Populations in a Cystic Fibrosis Patient Battista, J.R., *ed. PLoS ONE*, 8, e60225, 10.1371/journal.pone.0060225.

Yao, R., Burr, D.H. & Guerry, P. 1997. CheY-mediated modulation of *Campylobacter jejuni* virulence. *Molecular microbiology*, 23, 1021–1031, DOI 10.1046/j.1365-2958.1997.2861650.x.

Yogasundram, K., Shane, S.M. & Harrington, K.S. 2013. Prevalence of *Campylobacter jejuni* in selected domestic and wild birds in Louisiana. *Avian diseases*, 33, 664–667, Doi 10.2307/1591142.

Young, B.C., Golubchik, T., Batty, E.M., Fung, R., Larner-Svensson, H., Votintseva, A.A., Miller, R.R. 2012. Evolutionary dynamics of Staphylococcus aureus during progression from carriage to disease. *Proceedings of the National Academy of Sciences*, 109, 4550–4555, 10.1073/pnas.1113219109.

Young, C.R., Harvey, R., Anderson, R., Nisbet, D. & Stanker, L.H. 2000. Enteric colonisation following natural exposure to *Campylobacter* in pigs. *Research in Veterinary Science*, 68, 75–78, DOI: 10.1053/rvsc.1999.0335.

Young, K.T., Davis, L.M. & Dirita, V.J. 2007. *Campylobacter jejuni*: molecular biology and pathogenesis. *Nature reviews. Microbiology*, 5, 665–679, 10.1038/nrmicro1718.

Zautner, A.E., Malik Tareen, A., Groß, U. & Lugert, R. 2012. Chemotaxis in *Campylobacter Jejuni*. *European Journal of Microbiology and Immunology*, 2, 24–31, 10.1556/EuJMI.2.2012.1.5.

Zhang, G., Hubalewska, M. & Ignatova, Z. 2009. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nature Structural & Molecular Biology*, 16, 274–280, 10.1038/nsmb.1554.

Zhang, Q., Meitzler, J.C., Huang, S. & Morishita, T. 2000. Sequence polymorphism, predicted secondary structures, and surface-exposed conformational epitopes of *Campylobacter* major outer membrane protein. *Infection and Immunity*, 68, 5679–5689, 10.1128/IAI.68.10.5679-5689.2000.

Appendix

9.1. Accession Numbers of 16S rRNA Genes

Table 9.1 - Strain and Genbank accession numbers for 16S rRNA genes used in the generation of Figure 1.2.

Species	Strain	16S rRNA gene
A. butzleri	ATCC 49616	AY621116.1
C. fetus subsp. fetus	ATCC 27374T	DQ174127
C. fetus subsp. venerealis	NCTC 10354 T	DQ174131
C. coli	LMG 6440 T	AF372092
C. jejuni subsp. jejuni	NCTC 11351 T	AF372091
C. jejuni subsp. doylei	LMG 8843 T	DQ174144
C. lari	NCTC 11352 T	DQ174145
C. sputorum	LMG 7795 T	DQ174149
C. showae	CCUG 30254 T	DQ174155
C. upsaliensis	CCUG 14913 T	DQ174157
C. helveticus	NCTC 12470 T	DQ174161
C. curvus	ATCC 35224 T	DQ174165
C. concisus	ATCC 33237 T	DQ174166
C. gracilis	ATCC 33236 T	DQ174168
C. rectus	CCUG 20446 BT	DQ174169
C. mucosalis	ATCC 43264 T	DQ174173
C. hyointestinalis subsp. hyointestinalis	NADC 2006	M65009.1
C. hyointestinalis subsp. lawsonii	CCUG27631	AF097683.1
C. lanienae	NCTC 13004 T	DQ174181
C. hominis	NCTC 13146 T	DQ174182
C. insulaenigrae	NCTC 12927 T	DQ174183
C. avium	86/06	EU623473
C. subantarcticus	LMG 24377	NR_115081
C. peloridis	2314BVA	NR_042684
C. cuniculorum	150B	DQ400345
C. canadensis	L266	EF621894
C. volucris	LMG 24380	NR_116923
C. lari subsp. concheus	2897R	NR_042683

* - Genbank accession numbers.

9.2. *Campylobacter spp.* incidence at Queen's Medical Centre 2012-2014.

Table 9.2 - *Campylobacter spp.* incidence at Queen's Medical Centre, 2012-2014.

	Cases Per Year		
Month	2012	2013	2014
Jan	48	50	39
Feb	40	47	66
Mar	38	58	57
Apr	57	48	54
May	82	70	50
Jun	119	89	103
Jul	116	103	78
Aug	87	87	63
Sep	89	104	90
Oct	95	75	68
Nov	51	54	49
Dec	28	37	42
Total	**850**	**822**	**759**

9.3. Autospades.py

```
#!/usr/bin/env python
import os
import glob
import subprocess
print ''
print ''
print '######################'
print 'Welcome to AutoSPAdes'
print '######################'
print ''
print ''
# User input
directory = raw_input("Path to directory containing FastQ reads: ")
outputpath = raw_input("Path to output Fasta assemblies in: ")
print 'Assemblies will be collectively copied into your output path with logical
filenames'
print 'Do you want to remove assembly subdirectories after use?'
choice = raw_input("Y/N: ").lower()
print 'Use the careful flag?'
careful = raw_input("Y/N: ").lower()
if not os.path.exists(outputpath):
    os.makedirs(outputpath)
# List comprehension
r1files = list(glob.glob(os.path.join(directory,'*_R1_*')))
r1files.sort()
r2files = list(glob.glob(os.path.join(directory, '*_R2_*')))
r2files.sort()
yes = set(['yes','y','ye'])
no = set(['no','n',''])
rawname = [x.split(directory)[1].split('_')[0] for x in r1files]
rawname.sort()
subdirectoriesraw = [directory + x + '_' for x in rawname]
subdirectoriesraw.sort()
subdirectories = [x.split('_')[0] + '/' for x in subdirectoriesraw]
fastadirectories = [x + 'scaffolds.fasta' for x in subdirectories]
fastanames = [x + '.fasta' for x in rawname]
fastaoutput = [outputpath + x for x in fastanames]
# Spades invocation
if careful in yes:
    for opt1, opt2, opt3, in zip(r1files, r2files,subdirectories):
```

```
        subprocess.call(['spades.py', '--careful', '--pe1-1', opt1, '--pe1-2',
opt2, '-o', opt3])
if careful in no:
    for opt1, opt2, opt3, in zip(r1files, r2files,subdirectories):
        subprocess.call(['spades.py', '--pe1-1', opt1, '--pe1-2', opt2, '-o',
opt3])
# FASTA organisation (rename and collect)
for var1, var2, in zip(fastadirectories, fastaoutput):
    subprocess.call(['cp', var1, var2])
# End
if choice in yes:
    for remove, in zip(subdirectories):
        subprocess.check_call(['rm','-r', remove])
    print ''
    print ''
    print ''
    print '#'
    print '##'
    print '###'
    print '####'
    print '#####'
    print '######'
    print '#######'
    print '########'
    print '#########'
    print '##########'
    print '###########'
    print '############'
    print '#############'
    print '##############'
    print '###############'
    print 'Job Completed!'
    print '###############'
    print ''
    print ''
    print 'Subdirectories scrubbed.'
    print ''
    print ''
    print 'Fasta files located in:',
    print outputpath
    print ''
    print ''
    print 'FastQ files unzippd and stored in: ',
```

```
        print directory
        print ''
        print ''
        print ''
        print ''
        print 'Goodbye!'
        print '###############'
    if choice in no:
        print ''
        print ''
        print ''
        print '#'
        print '##'
        print '###'
        print '####'
        print '#####'
        print '######'
        print '#######'
        print '########'
        print '#########'
        print '##########'
        print '###########'
        print '############'
        print '#############'
        print '##############'
        print '###############'
        print 'Job Completed!'
        print '###############'
        print ''
        print ''
        print 'Fasta subdirectories and FastQ reads located in: ',
        print directory
        print ''
        print ''
        print 'Fasta files collected and stored in: ',
        print outputpath
        print ''
        print ''
        print ''
        print ''
        print 'Goodbye!'
        print '###############'
```

9.4. Autoprokka.py

```
#!/usr/bin/env python
import os
import subprocess
import glob
import shutil


directory = raw_input("Path containing FASTA assemblies: ")
outputraw = raw_input("Path to output annotation directories: ")
if not os.path.exists(outputraw):
    os.makedirs(outputraw)
fasta = list(glob.glob(os.path.join(directory,'*.fasta')))
output = [x.replace(directory, outputraw).replace('.fasta', '/') for x in fasta]
prefixis = [x.replace(directory, '').replace('.fasta', '') for x in fasta]


print ''
print ''
print 'Do you want to copy .gff into a single directory?'
print ''
choice = raw_input("Y/N? :").lower()
positive = set(['yes', 'ye', 'y', 'ys', ''])
if choice in positive:
    print ''
    print 'Which directory shall I place them in?',
    print ''
    outputgff = raw_input('Path to directory:')
    if not os.path.exists(outputgff):
        os.makedirs(outputgff)


for fastain, fastaout, pre in zip(fasta, output, prefixis):
        subprocess.call(['prokka', fastain, '-o', fastaout, '--prefix', pre])


if choice in positive:
    gffnames = [x.replace(directory, '').replace('.fasta', '.gff') for x in fasta]
    gfflocations = [''.join(x) for x in zip(output, gffnames)]
    gffoutput = [outputgff + x for x in gffnames]
    for gffin, gffout in zip(gfflocations, gffoutput):
```

```
        try:
            shutil.copyfile(gffin, gffout)
        except IOError as e:
            print 'The following files were not found, indicating that prokka did
not finish correctly.'
            print e
            print 'Please check FASTA input.'


print ''
print ''
print ''
print '########'
print 'Finished'
print '########'
```