



**Application of whole genome sequencing to fully characterise
Campylobacter isolates from the UK Infectious Intestinal Disease 1 and 2
Studies**

Project Code: FS101072

Funder: UK Food Standards Agency

Contractor: The University of Liverpool

Report prepared by

Craig Winstanley, Sam Haldenby, Christina Bronowski, Charlotte Nelson, Christiane Hertz-
Fowler, John Kenny, Roy Chaudhuri, Sarah O'Brien, Nicola Williams & Neil Hall

12th August 2015

Project Lead Contractor

Craig Winstanley

University of Liverpool

© Crown Copyright 2015

This report has been produced by the University of Liverpool under a contract placed by the Food Standards Agency (the FSA). The views expressed herein are not necessarily those of the FSA. The University of Liverpool warrants that all reasonable skill and care has been used in preparing this report. Notwithstanding this warranty, the University of Liverpool shall not be under any liability for loss of profit, business, revenues or any special indirect or consequential damage of any nature whatsoever or loss of anticipated saving or for any increased costs sustained by the client or his or her servants or agents arising in any way whether directly or indirectly as a result of reliance on this report or of any error or defect in this report.

1. Summary

Campylobacter is the most common cause of acute bacterial gastroenteritis worldwide. In the UK alone it causes an estimated 500,000 infections each year. There have been two large studies of Infectious Intestinal Disease in the UK community (IID1 in the mid 1990s and IID2 in 2008-2009). In both studies, *Campylobacter* was identified as the most common bacterial pathogen amongst patients presenting to primary care. Although there was little variation in the burden of illness between the two studies, the molecular epidemiology of the *Campylobacter* isolates from these studies has not been investigated previously.

The aim of this study was to use whole genome sequencing (WGS) and multilocus sequence typing (MLST) comparative analyses to characterise and compare isolates from the two survey periods, and to generate data that can be used to identify the potential sources for transmission of the pathogen to humans, by comparison with isolates from the environment, wildlife and farm animals.

WGS of all available *Campylobacter* isolates from the IID1 and IID2 studies was carried out using the Illumina platform. From the 504 samples received, WGS data was obtained for 470 *Campylobacter* isolates, comprising 351 from IID1 and 119 from IID2. Of these 416 were *C. jejuni* and 46 were *C. coli*. We also obtained WGS data from five *C. upsaliensis*, one *C. fetus*, one *Arcobacter butzleri* and one *Arcobacter* spp.

Analysis of MLST data extracted from the WGS data indicated that the most common clonal complexes found amongst the IID1 and IID2 strains reflected their abundance amongst the wider *Campylobacter* population. There were no clear variations of note between the IID1 and IID2 isolate collections with respect to breakdown according to MLST clonal complex (CC). The use of Single Nucleotide Polymorphism (SNP) phylogeny to cluster *C. jejuni* genomes based on either ribosomal loci (rMLST), or larger sets of core genes, confirmed the broad distribution of IID1 and IID2 isolates amongst the wider *Campylobacter* population, but highlighted some sub-divisions within MLST-based clonal complexes, and some small clusters specific to either IID1 or IID2. SNP-based phylogenetic analysis of *C. coli* isolates from IID1 and IID2 indicated that they all cluster within Clade 3, a clade associated previously with clinical isolates and agricultural sources.

Using PacBio sequencing, we were able to add a further 17 high quality reference genomes to the general database, eleven of which (ten *C. jejuni* and one *C. coli*) assembled as a single chromosome. From 14 high quality PacBio genomes and three previously available complete genomes, we defined a *C. jejuni* core genome of 1261 genes.

These data provide an excellent baseline for monitoring shifts in the UK population of *Campylobacter* associated with gastrointestinal infections. By combining survey data of this nature with analyses of other isolate collections from non-human sources, it will be possible to identify changing trends and shifts in the relative importance of potential sources of transmission.

2. Executive Summary

Campylobacter is the most common cause of acute bacterial gastroenteritis worldwide. Although most cases are attributable to infection via poultry, there is evidence that other agricultural or environmental sources can play a role. Different strain types of *Campylobacter* are associated with the different potential infection sources. This has been shown using the relatively crude method of multilocus sequence typing (MLST), which examines the sequences of a small number of genes for each strain, and categorises them accordingly, grouping related strains as “clonal complexes” (CC). In recent years, new and more affordable technology for whole genome sequencing (WGS), whereby the entire genetic content of a strain can be analysed, has become available, offering the potential for comparisons with a much higher resolution than before.

The overall aim of this study was to characterise UK *Campylobacter* strains associated with human campylobacteriosis by WGS of isolates from two infectious disease surveys (IID1 and IID2) carried out during two different time periods (IID1 [1993 – 1996] and IID2 [2008 – 2009]). Such WGS data can then be used to identify markers to assist with source attribution by integration of the data from IID study strains with published data obtained from non-human sources and genome data being derived from current funded projects at Liverpool.

The specific objectives of the study were to:

- (i) Use WGS on the Illumina sequencing platform, to obtain a comprehensive genome sequence data-set from the IID isolates, submitted to a publically accessible database
- (ii) Analyse MLST data from the IID isolates to place the collection in the context of previous studies based on MLST
- (iii) Carry out genome-wide phylogenetic analysis (study of evolutionary relationships) of the IID strains compared to others available in the wider database
- (iv) Use higher quality WGS on the PacBio sequencing platform to provide additional *Campylobacter* reference genomes for exploitation by the research community

Campylobacter isolates from the IID1 and IID2 studies were obtained from Public Health England (PHE) and stored in Liverpool. Genomic DNA was extracted from bacterial isolates using either (i) the QIAamp kit (Qiagen) or (ii) the Wizard Genomic DNA Purification Kit (Promega). In this project, we found that the Qiagen kit extracts yielded the best sequencing

results. For all of the isolates from which DNA of sufficient quality and quantity could be obtained, libraries were made in preparation for sequencing. Having tested the Nextera XT DNA Sample Prep Kit on a sub-set of samples, we used the TruSeq Nano DNA Sample Preparation Kit for library preparation for the majority of the strains. Following sequencing using the Illumina HiSeq platform, the data obtained from isolates were automatically analysed through an in-house bioinformatics pipeline, Campype, enabling (i) extraction of MLST data, (ii) multi-sequence alignment, (iii) phylogenetic analysis based on ribosomal genes (rMLST) or a set of 599 core genes (shared by all of the strains).

For a sub-set of 26 isolates, we obtained additional quantities of genomic DNA and libraries were made for sequencing on the PacBio platform according to the manufacturer's instructions. In order to obtain sufficient DNA, it was necessary to repeat extractions up to 20 times, which makes this approach far more labour intensive. After some difficulties in obtaining sequence data, an additional wash was performed on the libraries using the Magbead clean-up protocol (PacBio). This resolved the main problem, enabling us to obtain PacBio sequence data for 16 of the isolates.

From an initial 504 samples received, WGS data was obtained for 470 *Campylobacter* isolates using the Illumina platform; comprising 351 from IID1 and 119 from IID2. Of these 416 were *C. jejuni* and 46 were *C. coli*. We also obtained WGS data from five *C. upsaliensis*, one *C. fetus*, one *Arcobacter butzleri* and one *Arcobacter* spp.

Analysis of MLST data extracted from the WGS data, coupled to comparisons to large publically accessible databases from previous studies worldwide, indicated that the most common clonal complexes found amongst the IID1 and IID2 strains reflected their abundance amongst the wider *Campylobacter* population. There were no clear variations of note between the IID1 and IID2 isolate collections with respect to breakdown according to MLST clonal complex, suggesting that the strain types infecting humans have not changed significantly in over a decade.

Genome-wide phylogenetic analysis confirmed the broad distribution of IID1 and IID2 isolates amongst the wider *C. jejuni* population, but highlighted some discrepancies in comparison to clustering based on MLST. There were also some small clusters of closely-related strains specific to either IID1 or IID2. Phylogenetic analysis of *C. coli* isolates from

IID1 and IID2 indicated that they all cluster within Clade 3, a clade associated previously with clinical isolates and agricultural sources.

Using PacBio sequencing, we were able to add a further 17 high quality reference genomes to the general database, eleven of which (ten *C. jejuni* and one *C. coli*) assembled as a single chromosome. From 14 high quality PacBio genomes and three previously available complete genomes, we defined a *C. jejuni* core genome of 1261 genes.

Although there were no significant shifts in the overall populations of *Campylobacter* infecting humans between the two IID studies, some small clusters were specific to either IID1 or IID2, and are worthy of deeper analysis. WGS data enables high resolution and extremely accurate analysis of relationships between bacterial strains and elucidation of patterns of transmission. This would, for example, allow rapid linkage of strains that have been acquired in an outbreak or from a common source.

The data obtained in this study provides an excellent baseline for monitoring shifts in the UK population of *Campylobacter* associated with gastrointestinal infections, by comparison with data from future surveys. This would require regular monitoring using WGS analysis.

By combining survey data of this nature with analyses of other isolate collections from non-human sources, it will be possible to identify changing trends and shifts in the relative importance of potential sources of transmission.

Glossary

ATL:	ATL buffer is a lysis buffer in the Qiagen DNA extraction kit
CC:	clonal complex
CDS:	coding sequences
Clade:	an evolutionarily-related group containing a common ancestor and its descendants
Contigs:	contiguous sequences
GI:	GenInfo (GI) identifiers are assigned consecutively to each sequence record processed by the National Center for Biotechnology Information
GP:	General Practitioner
HGAP:	Hierarchical Genome Assembly Process
IID1:	Infectious Intestinal Disease Study 1 (1993 – 1996)
IID2:	Infectious Intestinal Disease Study 2 (2008 – 2009)
LS-BSR:	Large Scale Blast Score Ratio, a piece of software for comparing the genetic content of hundreds/thousands of bacterial genomes
MLST:	Multilocus sequence typing
L50:	a statistic relating to the length of contiguous sequences
N50:	a statistic relating to the length of contiguous sequences
NCBI:	National Center for Biotechnology Information
ORF:	open reading frame (predicted gene)
PacBio:	Pacific Biosciences sequencing platform
PBS:	phosphate buffered saline
PCR:	polymerase chain reaction (amplification)
PGFE:	pulsed-field gel electrophoresis
PHE:	Public Health England
QC:	quality control
qPCR:	quantitative polymerase chain reaction (amplification)
Qubit:	a DNA quantification system based on fluorometry
RAPD:	random amplified polymorphic DNA
rMLST:	ribosomal multilocus sequence typing
SDW:	sterile distilled water
SNP:	single nucleotide polymorphism
ST:	sequence type
VCF:	variant call format

VFG: virulence factor genes

WGS: whole genome sequencing

3. Full Report

3.1 Background

Campylobacter is the most common cause of acute bacterial gastroenteritis worldwide. In the UK alone it causes an estimated 500,000 infections each year (1) and presents an economic burden of over £1 billion per annum (2). Campylobacteriosis typically lasts for about a week, is often characterised by bloody diarrhoea, cramping, abdominal pain and fever, and may be accompanied by nausea and vomiting. Occasionally, in immunocompromised patients, the pathogen can cross into the bloodstream, leading to much more serious systemic infections. This human pathogen is also a major predisposing cause of the peripheral nervous system disorder, Guillain-Barré Syndrome (3). Although the majority of cases are caused by *C. jejuni* (approximately 90%), *C. coli* is also important, contributing to most of the remaining cases. Both species can be isolated from a wide variety of sources, ranging from human infection cases, through farm animals, to wild birds and animals and the environment.

Two large scale prospective studies of infectious intestinal disease (IID) incidence and aetiology have been carried out in the United Kingdom, named Infectious Intestinal Disease Studies 1 and 2 (IID1 and IID2, respectively). IID1 was carried out between 1993 – 1996 and IID2 was carried out between 2008 – 2009 (1, 4). Both included a community cohort and a healthcare presentation study. Faecal samples were examined for a range of bacterial, viral and protozoal pathogens. *Campylobacter* was the most common bacterial pathogen among both the community cohort and cases presenting to healthcare, though prevalence was much greater amongst cases presenting to GPs (1).

There have been a number of approaches used to sub-divide species of *Campylobacter*; especially *C. jejuni* and *C. coli*, including macro-restriction pulsed-field gel electrophoresis (PFGE) (5), flagellin genotyping (6), random amplified polymorphic DNA (RAPD) typing (7) and ribotyping (8). However, the development of a multilocus sequence typing (MLST) scheme for *Campylobacter* was a significant step forward in the study of diversity amongst *Campylobacter* populations and the relationships between species within the genus (9). MLST enables unequivocal data to be compared between laboratories world-wide through

the use of a readily accessible database (pubmlst.org/campylobacter), which contains data for > 31000 isolates.

The initial MLST scheme was based on the analysis of sequences from seven housekeeping genes (*aspA*, *glnA*, *gltA*, *glyA*, *pgm*, *tkt* and *uncA*) and allows the assignment of isolates to clonal complexes (clusters of closely-related sequence types). Using this approach, it was possible to identify the most abundant common clonal complexes (such as ST-21), though it is also evident that the *C. jejuni* population overall is highly diverse (9, 10). Others have extended the MLST scheme for improved applicability to other *Campylobacter* species (11, 12).

It is interesting to note that most cases of campylobacteriosis occur as isolated, sporadic cases, rather than as part of larger outbreaks, as typically seen with other bacterial pathogens associated with diarrhoea. It is believed that zoonotic transmission of *Campylobacter* spp. to humans occurs primarily through the consumption and handling of livestock, with poultry being the most common source. However, it is clear that other infection routes, including the environment, may also contribute. For example, it has even been suggested that flies might be a direct source of infection of humans, as well as contributing to the spread of *Campylobacter* to farm animals (13).

Campylobacter has been isolated from diverse animal, human and environmental sources and the isolates obtained subjected to genotyping. Using molecular typing coupled with epidemiological analysis we are now in a position to identify and track specific strain types of *C. jejuni* and *C. coli*. Several studies have sought to determine the prevalence of specific clones amongst *C. jejuni* isolates from diverse sources by applying MLST (9, 14-22). These studies show that whilst some MLST clonal complexes, such as the ST-21 complex, are widespread, others, such as the ST-61 complex, have a more restricted distribution. Although generally considered to be poor survivors outside of their animal hosts, some *C. jejuni* appear to be more able to survive and persist in environmental niches (18, 23, 24). For example, in a study of *C. jejuni* in a specific area of cattle farmland in the UK (18), isolates from the ST-45 complex were much more frequently isolated from environmental water sources, than were other common clonal complexes. Thus by determining the prevalence of specific strain types amongst isolates from potential sources, and comparing with similar data from isolates associated with infections in humans, we are able to model the relative contributions of

particular sources to transmission (21, 22, 25). This has considerably enhanced our understanding of transmission by allowing an insight into the relative contributions of various sources to human campylobacteriosis cases. This genetic attribution approach has demonstrated that the majority of human infections are associated with retail chicken meat (21, 22).

Ribosomal MLST (rMLST) (26) is an alternative approach for studying relationships between bacterial strains of the same species that has been applied to *Campylobacter* (27). rMLST uses the sequence data from 53 genes encoding the bacterial ribosome protein subunits.

Campylobacter are spiral members of the Epsilonproteobacteria with small genomes (typically 1.5 – 2 Mb). Although MLST has proved to be an excellent tool for the study of *Campylobacter* populations, the advent of affordable whole genome sequencing (WGS) technologies means that a scheme based on much wider genomic comparisons is likely to supersede MLST. This presents considerable opportunities to improve our understanding of both the epidemiology and the fundamental biology of this pathogen. Since the first genome sequence (of strain NCTC11168) was published in 2000 (28), numerous other *Campylobacter* genomes have been sequenced, revealing extensive within-species diversity (29-31). Building on these comparative studies involving relatively small numbers of isolates, WGS analysis has since been used in genome-wide association studies to identify *Campylobacter* genes associated with isolates from a particular host (32), and to develop more detailed and sophisticated sequence-based typing approaches (33).

Although a number of previous studies have used comparative genomic hybridisation (microarrays) to characterise genomic variations between strains from multiple sources (20, 30, 34-37), identifying variable genomic regions, especially in loci such as those encoding lipooligosaccharide (LOS), capsule or restriction-modification (RM) systems, this technique is restricted to analysis of those genes represented on the array, and cannot identify minor variations within genes. WGS offers superior resolution and also aids identification of novel genomic regions not already represented in the database.

In parallel with the shift from MLST to WGS, accessible and interactive databases have been developed to enable comparative analysis of new WGS data with previous genomes (38),

notably the Bacterial Isolate Genome Sequence Database (BIGSDB; <http://pubmlst.org/software/database/bigfdb/>).

3.2 Aims of this study

The overall aims of this study were:

- (i) to characterise UK *Campylobacter* strains associated with human campylobacteriosis by WGS; and
- (ii) to provide a dataset to assist with source attribution by integration of the data from IID study strains with published data obtained from non-human sources and genome data being derived from current funded projects at Liverpool.

3.3 Outcomes of the study

The envisaged outcomes of the study were:

- (i) A comprehensive genome sequence data-set from the IID isolates, submitted to a publically accessible database.
- (ii) Analysis of MLST data from the IID isolates to place the collection in the context of previous studies based on MLST.
- (iii) Genome-wide phylogenetic analysis of the IID strains compared to others available in the wider database.

In addition, by sequencing this strain collection we will considerably enhance the community's knowledge on what constitutes the core genome of *Campylobacter*, especially in relation to isolates associated with human infections, with the potential to link variations between strains (either in accessory genome content or in SNP variations within the core genome) with other factors such as putative source or, potentially, clinical severity, as well as other important phenotypes, such as survival characteristics in the environment or during food processing.

3.4. Experimental Procedures

3.4.1 Putative *Campylobacter* isolates

A total of 504 IID samples were retrieved from PHE and sub-cultured immediately on Columbia blood agar (5% v/v blood) incubated at 37°C in microaerobic conditions

(CampyGen, Oxoid). Following growth, multiple aliquots of each isolate were stored at -85°C.

3.4.2 DNA extractions for Illumina sequencing

For the Qiagen DNA extraction method, bacteria were retrieved from frozen storage and cultured on Columbia blood agar incubated at 37°C in microaerobic conditions (CampyGen, Oxoid) for 48 h. One or a few uniform colonies were used to inoculate 10-15ml of Mueller Hinton broth (with *Campylobacter* growth supplement, Oxoid) in a tissue culture flask with a vented lid. Cultures were incubated for 24 h with gentle shaking under microaerobic conditions at 37°C, after which cells were harvested by centrifugation (3000 x g for 10 min) in a Falcon tube. After removal of the supernatant, the cells were re-suspended in 1ml PBS and transferred to 1.5ml tubes before further centrifugation for 3 min at 10,000 x g. The supernatant was removed and the pellet was used for extraction using the Qiagen method, as described below.

For the Promega Wizard Preparation method, bacteria were cultured on Columbia blood agar incubated at 37°C in microaerobic conditions (CampyGen, Oxoid) for 48 h. A sub-culture was then carried out on the same media for 24 h. Thick bacterial suspensions were made directly from these culture plates in the Nuclei Lysis Solution (Promega) and DNA extracted using the Promega method as described below.

3.4.3 DNA extraction methods

3.4.3.1 Qiagen method

Once the bacterial pellets had been re-suspended in 180 µl ATL buffer, DNA extractions were carried out using the QIAamp kit, following the manufacturer's protocol. The optional treatment with ribonuclease was included. At the final stage, elution was carried out in 200 µl SDW, incubating the column for 5 min, after which there was a second elution carried out, in a further 200 µl of SDW with incubation for 5 min, to ensure maximum yields.

3.4.3.2 Promega method

DNA was extracted from cell suspensions of bacterial cultures grown on solid media, using the Wizard Genomic DNA Purification Kit (Promega) and the manufacturer's protocol for Gram-negative bacteria.

3.4.4 Library preparation for Illumina sequencing

3.4.4.1 Nextera XT

Initially the Nextera XT DNA Sample Prep Kit (Illumina) was chosen due to both the low input material requirements of this kit and the low cost. One ng of material was used as input material for the Nextera XT DNA Sample Prep Kit. Following 12 cycles of amplification, the libraries were purified using 0.6x AMPure XP beads (Beckman Coulter, Inc.). Each library was quantified using the Qubit (Life Technologies) and the size distribution assessed using the Agilent 2100 Bioanalyzer. The samples were then pooled and the final library assessed using the Agilent 2100 Bioanalyzer. We saw around a 20% failure rate in the sequencing of the initial batch of 50 samples using this library preparation kit. It was therefore decided to switch to the TruSeq Nano DNA Prep Kit (Illumina) as an alternative (below).

3.4.4.2 TruSeq Nano

Two hundred ng of material was used as input material for the TruSeq Nano DNA Sample Preparation Kit. The material was sheared using the Covaris S2 ultra-sonicator following the 550 bp insert size protocol. Due to budget constraints only half (100ng) of the sheared material was cleaned using 1.6x Sample Purification beads, and half volumes of all reagents were used throughout the entire protocol. Samples were prepared in a 96-well plate format. The samples were size-selected using the sample purification beads as directed in the protocol. Following 8 cycles of amplification the libraries were purified using Sample Purification beads. Each library was quantified using Qubit and the size distribution assessed using the Agilent 2100 Bioanalyzer. The samples were then pooled and the final library assessed using the Agilent 2100 Bioanalyzer.

The quantity and quality of the TruSeq Nano DNA library pools were assessed using the Qubit and Bioanalyzer, and subsequently subjected to quantitative PCR (qPCR) using the Illumina Library Quantification Kit from Kapa on a Roche Light Cycler LC480II, according to manufacturer's instructions.

3.4.5 Library preparation for PacBio sequencing

Libraries were made for PacBio (Pacific Biosciences) sequencing according to the manufacturer's instructions. Genomic DNA was sheared with a Covaris g tube and cleaned using AMPure XP beads. This was exonuclease treated, end repaired, AMPure cleaned and ligated to the adapter. SMRT cell templates were recovered following a further exonuclease treatment and AMPure clean up. Sizes were checked on a Fragment analyser and the quantity checked using Qubit High sensitivity kit.

Initially, samples were annealed to sequencing primer and bound to DNA polymerase using the quantities defined by the PacBio binding calculator. These annealed complexes were bound to Magbeads and sequenced using P6 polymerase and C4 chemistry. However, loading was very poor and little sequence data were obtained. Following discussions with PacBio technical support, it was suggested that there was inhibition of the polymerase which was sample specific and was not removed by the AMPure beads. This could be related to the DNA extraction method or the source of the material used for extraction.

An additional wash was performed on the libraries using the Magbead clean-up protocol. This method recovers material which is only compatible with the primer. Essentially all the library was annealed to primer and this was subsequently bound to washed Magbeads. After binding, the bound DNA was washed and recovered. After measuring the new concentration, the cleaned sample was reannealed to the primer and bound to DNA polymerase according to the binding calculator values. This complex was bound to Magbeads and sequenced with P6 polymerase and C4 chemistry. The Magbead-cleaned DNA yielded between 500-1000 Mb per SMRT cell.

3.4.6 Sequencing on the Illumina platform

Genomes were sequenced by paired-end sequencing (2 x 100bp) of 96 indexed samples per run, multiplexed into one lane of the Illumina HiSeq platform and generating in excess of 240Mb mappable reads per lane.

3.4.7 Bioinformatic analysis for Illumina-derived sequence data

Sequence data obtained from isolates were automatically analysed through an in-house pipeline, Campype (Figure 1). As input, Campype takes sequence-adaptor trimmed paired-end reads in FASTQ format. Reads are trimmed to remove low quality bases using Sickle 1.210 (<https://github.com/najoshi/sickle>) with a minimum window quality score of 25.

Trimmed reads are then assembled into contigs and subsequently scaffolded using SPAdes (39).

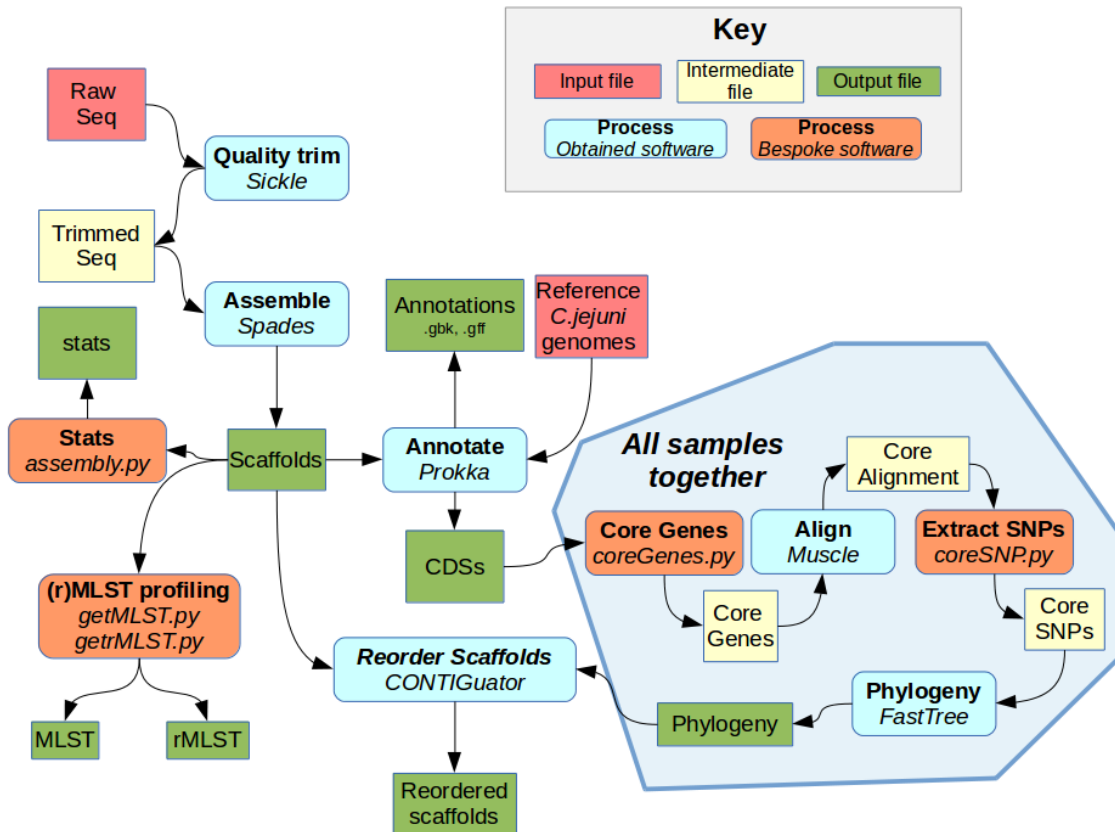
MLST types (using the 7 loci scheme: *aspA*, *glnA*, *gltA*, *glyA*, *pgm*, *tkt*, *uncA*) were extracted from the assembled scaffolds by aligning them against a database of known alleles with LAST (40). Similarly, rMLST types (using the 65 loci scheme: BACT000001-BACT000065) were extracted from scaffolds by aligning them against a database of known rMLST sequences using BLAST (41). Campype predicts ORFs within scaffolds and subsequently annotates scaffolds with Prokka (42).

For the *Campylobacter* SNP phylogeny, coding sequences (CDS) from each complete genome included in the analysis were identified and iteratively mapped to each other with BLAST. Only CDSs that were present in all samples were kept. For each CDS among the samples, multiple alignments were generated using MUSCLE (43). For each multiple alignment, the following were removed: (a) sites where gaps were present and (b) non-variable sites, leaving only sites which are variant in at least one sample. The SNP sites for each CDS were concatenated together for each species to generate one large alignment and a tree was generated using FastTree (44). The rMLST tree was constructed using concatenated sequence data from the ribosomal loci. The *Campylobacter* SNP phylogeny was used to classify strains by determining the closest completely sequenced strain on the tree.

To improve/finish the assembly, scaffolds were mapped to the closest reference sequence and reordered based on this mapping, using CONTIGuator (45). Finally, Campype calculates basic assembly statistics, such as total assembly size, N50 value and L50 value.

eBURST (46) was used to produce relationship clusters of *Campylobacter* isolates based on MLST profiles available at <http://eburst.mlst.net>.

Figure 1. Overview of the Campype pipeline.



3.4.8 SNP phylogeny combining Illumina data with other genomes in the database

Assemblies of 928 *C. jejuni* isolates (origin: UK, derived from humans) were downloaded from <http://pubmlst.org/databases/>. These assemblies and 414 Illumina assemblies from this study were used as input for Large scale Blast Score Ratio (LS-BSR) analysis (47). LS-BSR was used to determine the core and pan genome of all assemblies. The core genome was then extracted using the LS-BSR tool `extract_core_genome.py` which generated a sequence containing concatenated core genes for each sample.

Variant sites among all core genomes were extracted (33627 sites) using an in-house script and the output was used to generate a SNP phylogeny with FastTree (using the generalized time-reversible model).

3.4.9 Bioinformatics analysis for PacBio-derived sequence data.

Assemblies were carried out on the PacBio SMRT portal using HGAP (Hierarchical Genome Assembly Process; SMRTAnalysis version v2.2.0.p1), with provided reference genome guide size of 1.7 Mb. HGAP first preassembled reads by creating a set of seed reads, representing the longest portion of the read length distribution, and mapping single-pass reads against them using BLASR (48).

From these mappings, a highly accurate consensus sequence was generated (preassembled contigs) with pbdagcon (<https://github.com/PacificBiosciences/pbdagcon>). Prior to assembly with the Celera assembler, preassembled contigs underwent quality trimming. The assembly was carried out with the Celera assembler (49), within HGAP.

Resulting assemblies were sorted from longest to shortest contig size, and the top ten contigs for each assembly were used as queries for BLAST searches against the non-redundant database (nt), to determine the sequence origins.

3.4.10 *C. jejuni* core genome analysis from PacBio data

Scaffolds from all PacBio genome assemblies were annotated using BLAST (41) to determine the species from which the sequences were derived. All assemblies that were not of a *C. jejuni* genome were removed. Also, all non-*jejuni* scaffolds from *C. jejuni* assemblies were removed. The remaining assemblies, along with the existing *C. jejuni* reference complete genomes for isolates 81-176, RM1221 and NCTC11168, were used as input for LS-BSR (47). LS-BSR was used to determine the core and pan-genome of all assemblies. The core genome was then extracted using the LS-BSR tool `extract_core_genome.py` which generated a sequence containing concatenated core genes for each sample.

Variant sites among all core genomes were extracted (107760 sites) using an in-house script and the output was used to generate a SNP phylogeny with FastTree (using the generalized time-reversible model).

Additionally, all genes in the pan-genome were annotated using BLAST (41) (`blastx` vs all bacterial proteins in the NCBI database [www.ncbi.nlm.nih.gov]; minimum e-value of $1e^{-50}$). Determined GI identifiers were used as a look-up to obtain further annotation information from UniProt (www.uniprot.org). A table showing the relatedness of all peptides in the pan-genome across all genomes was generated using the BSR matrix output of LS-BSR and

this annotation information.

3.4.11 Variations in carriage of genes implicated in virulence

To investigate the carriage of genes implicated in pathogenicity, DNA sequences of known virulence factor genes (VFGs) were obtained from the Virulence Factor Database: VFDB (50) (<http://www.mgc.ac.cn/VFs/Down/VFs.ffn.gz>). These sequences were used as an initial query for BLASTN searches (41) (e-val 1e-50) against the contigs of each sample. Resulting best matches were extracted from the contigs and used as a BLAST query against VFGs (e-val 1e-50), to ensure best reciprocal matches. Scoring for the presence/absence of each factor was calculated as follows:

percentage identity of the reciprocated BLAST x (BLAST hit length / VFG length).

3.5. Results and Discussion

3.5.1 *Campylobacter* isolates

A total of 504 samples were retrieved from PHE, cultured and stored in Liverpool. These comprised 376 from IID1 and 128 from IID2. For some of the isolates, DNA was extracted using the Promega method. However, our initial sequencing failure rate was high. We therefore switched to the Qiagen DNA extraction method, and this significantly improved the proportion of successful sequencing runs. It is worth noting that we found it impossible to predict which samples would fail by using the QC data.

3.5.2 Illumina sequence data

For 38 of the isolates we were unable to either (i) grow the isolate from storage or (ii) extract DNA of sufficient quantity and quality to proceed. For the latter category, all isolates were subjected to DNA extraction on a minimum of two occasions. Following sequencing, some isolates were rejected on the basis that there were variations within their MLST loci (i.e. mixed MLST loci, suggesting the presence of more than one strain). After removing those genomes where mixtures were evident, WGS data was obtained for 466 isolates, including 412 *C. jejuni* and 46 *C. coli*. The data for these two dominant species comprised 341 from IID1 and 117 from IID2 (Table 1). A summary of the WGS data (MLST data, number of contigs, assembly length, N50 and L50 values) is shown in Appendix 1 (Further details of all appendices are presented in 3.10).

3.5.3 MLST analysis based on WGS data

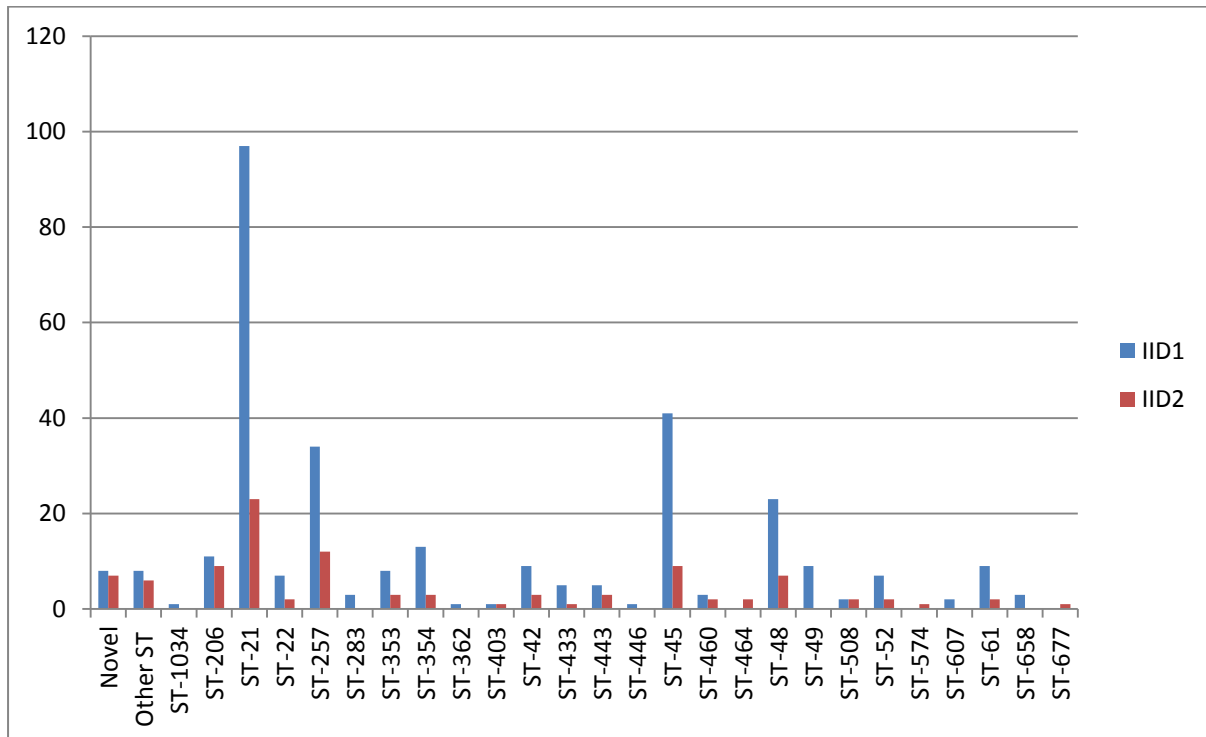
A summary of the MLST profiles extracted from WGS data obtained using the IID1 and IID2 isolates is shown in Table 1. This include the additional four isolate genomes sequenced using the PacBio system (not previously sequenced on Illumina platform), as referred to in section 3.5.8.1.

Table 1. Summary of MLST data.

MLST Clonal Complex	Total	IID1	IID2
<i>C. jejuni</i>			
Novel	15	8	7
Other ST	14	8	6
ST-1034	1	1	0
ST-206	20	11	9
ST-21	121	98	23
ST-22	9	7	2
ST-257	47	35	12
ST-283	3	3	0
ST-353	11	8	3
ST-354	16	13	3
ST-362	1	1	0
ST-403	2	1	1
ST-42	12	9	3
ST-433	6	5	1
ST-443	8	5	3
ST-446	1	1	0
ST-45	51	41	10
ST-460	5	3	2
ST-464	2	0	2
ST-48	31	23	8
ST-49	9	9	0
ST-508	4	2	2
ST-52	9	7	2
ST-574	1	0	1
ST-607	2	2	0
ST-61	11	9	2
ST-658	3	3	0
ST-677	1	0	1
<i>C. coli</i> ST-828	38	24	14
<i>C. coli</i> ST-1009	6	5	1
Other <i>C. coli</i>	2	1	1
Other <i>Campylobacter</i> or <i>Arcobacter</i>	8	8	0
TOTAL SEQUENCED	470	351	119
Total samples	504	376	128

Figure 2 shows the distribution of isolates amongst clonal complexes for the IID1 and IID2 *C. jejuni* isolates only. Figure 3 shows the clonal complex data plotted as a percentage of total *C. jejuni* for either IID1 or IID2.

Figure 2. Distribution of *C. jejuni* IID1 and IID2 isolates according to MLST clonal complex.



The most common clonal complexes identified were shared by both IID1 and IID2 collections, with the most prevalent being ST-21, ST-45, ST-257 and ST-48 for *C. jejuni* and ST-828 for *C. coli*. The MLST database (www.mlst.net) currently contains >31,000 *Campylobacter* submissions. Figure 4 shows in order the most abundant MLST clonal complexes in the database. Broadly speaking, the most common clonal complexes found amongst the IID1 and IID2 strains reflect their abundance amongst the wider *Campylobacter* population. There were no clear variations of note between the IID1 and IID2 collections with respect to breakdown according to MLST clonal complex (Figure 3).

Appendix 2 shows an eBURST figure based on the MLST profiles of the IID1 and IID2 isolates amongst the wider population in the MLST database. The IID1 and IID2 isolates

were widely distributed. An eBURST representation of just the IID1 and IID2 isolates is shown in Appendix 3.

There were 15 novel MLSTs identified for the first time in this study, and the proportion of novel STs was higher in IID2 than IID1 (Table 1; Figure 3). Data for the novel MLSTs have been submitted to the mlst.net database (Table 2).

Table 2. Novel MLSTs submitted to the mlst.net database.

Isolate		aspA	glnA	gltA	glyA	pgm	tkf	uncA	ST
C031641	IID1	47	75	5	5	2	48	8	7628
C031663	IID1	2	1	21	3	2	78	78	7629
C031671	IID1	8	10	2	6	11	12	6	7630
C031810	IID1	7	17	2	15	23	59	12	7631
C033063	IID1	8	30	2	2	89	59	6	7632
C033204	IID1	7	1	2	2	4	1	6	7633
C033241	IID1	183	7	10	4	1	7	1	7634
H083540698	IID2	18	89	77	109	122	269	16	7635
H083720447	IID2	47	55	5	10	89	48	8	7636
H090200185	IID2	277	105	20	143	134	15	16	7637
H090520098	IID2	2	75	4	531	141	34	1	7638
H090620069	IID2	22	7	10	4	42	51	1	7639
H090840324	IID2	2	21	5	37	499	1	5	7640
H092960272	IID2	2	1	12	3	340	1	5	7641
C031633	IID1	6	114	6	2	2	1	1	7642

3.5.4 Analysis of IID1 and IID2 WGS data using rMLST.

A phylogenetic tree constructed of IID1 and IID2 isolates based on rMLST loci is shown in Appendix 4. (IID1 isolates, blue; IID2 isolates, red).

Although isolates from the two datasets are widely distributed throughout the tree, small clusters specific to either IID1 or IID2 strains were apparent. Closer analysis of the metadata in the future should reveal further information about the biological basis of this clustering.

Figure 3. Distribution of isolates in each clonal complex as a percentage of total *C. jejuni* for IID1 and IID2. Clonal complex groups are ranked according to abundance in IID1

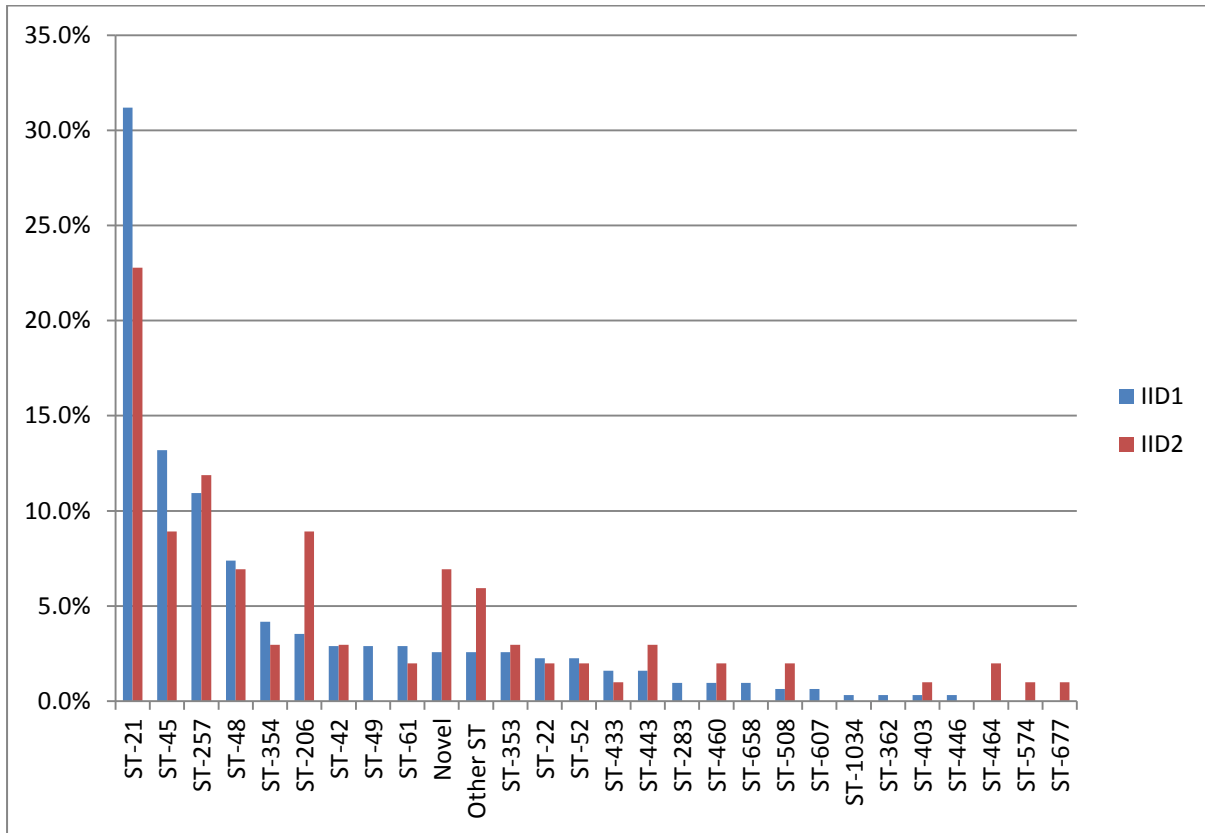
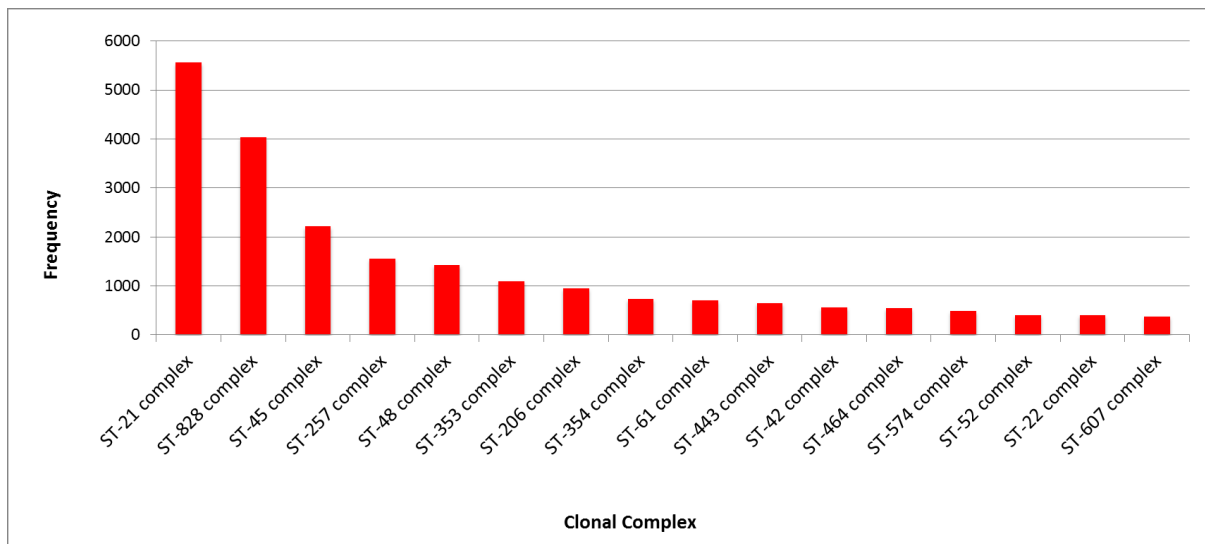


Figure 4. The most abundant *Campylobacter* clonal complexes in the MLST database.



3.5.5 SNP phylogeny based on the core genome.

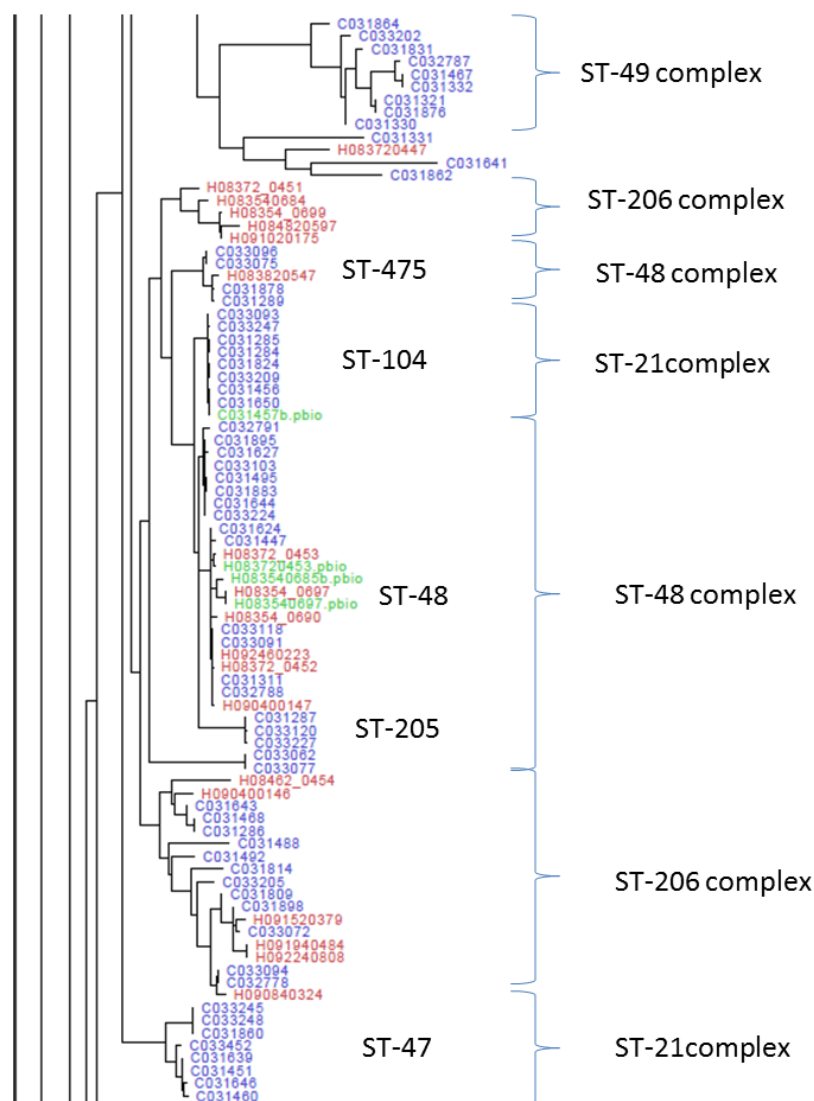
From the Illumina data it was possible to construct a phylogenetic tree for *C. jejuni* based on SNPs within a core genome set of 599 genes and comprising >19000 variable sites. Although this cannot be described as the complete *C. jejuni* core genome, it was the most complete dataset available to us whilst ensuring that all strains gave good quality sequence data for each of the genes included. The complete tree is shown in Appendix 5. Appendix 6 shows a more detailed breakdown of the tree and comparisons with MLST data (further explanation provided in 3.10 for appendices). It is clear that for some clonal complexes, different STs within the clonal complex do not necessarily cluster with each other. For example, Figure 5 shows how the ST-48 clonal complex is interspersed with STs from other clonal complexes (such as ST-21) when this wider dataset of genes is used. Likewise, it demonstrates how individual STs within the ST-21 clonal complex isolates do not cluster with each other. In this figure, examples of small clusters specific to IID1 or IID2 can also be seen. Closer analysis of the metadata can be used in the future to obtain further information about the biological basis for this kind of clustering.

The fact that MLST clonal complexes breakdown when more sophisticated analysis is carried out using WGS, suggests that conclusions based on MLST clonal complexes should be treated with some caution, and reflects the greater resolution of the WGS approach. However, it is worth noting that strains sharing the same individual ST did tend to cluster together.

The Illumina sequence data generated in this project were used in combination with 928 *C. jejuni* isolate genomes present in the pubmlst.org database. These comprised the majority of genomes in there, and were all UK isolates from humans. The phylogenetic tree derived is shown in Appendix 7 (IID1 isolates, blue; IID2 isolates, red). The tree confirms both (i) that the IID1 and IID2 isolates are widely distributed amongst the broader population and (ii) that there are small clusters specific to either IID1 or IID2. An example is shown in Figure 6.

Figure 5. Isolates within MLST clonal complexes do not all cluster together when core genome SNP phylogeny is used. Numerous small clusters of isolates specific to either IID1 (all blue) or IID2 (all red) were observed.

Key: IID1 isolates blue; IID2 isolates red (PacBio sequenced strains, green).

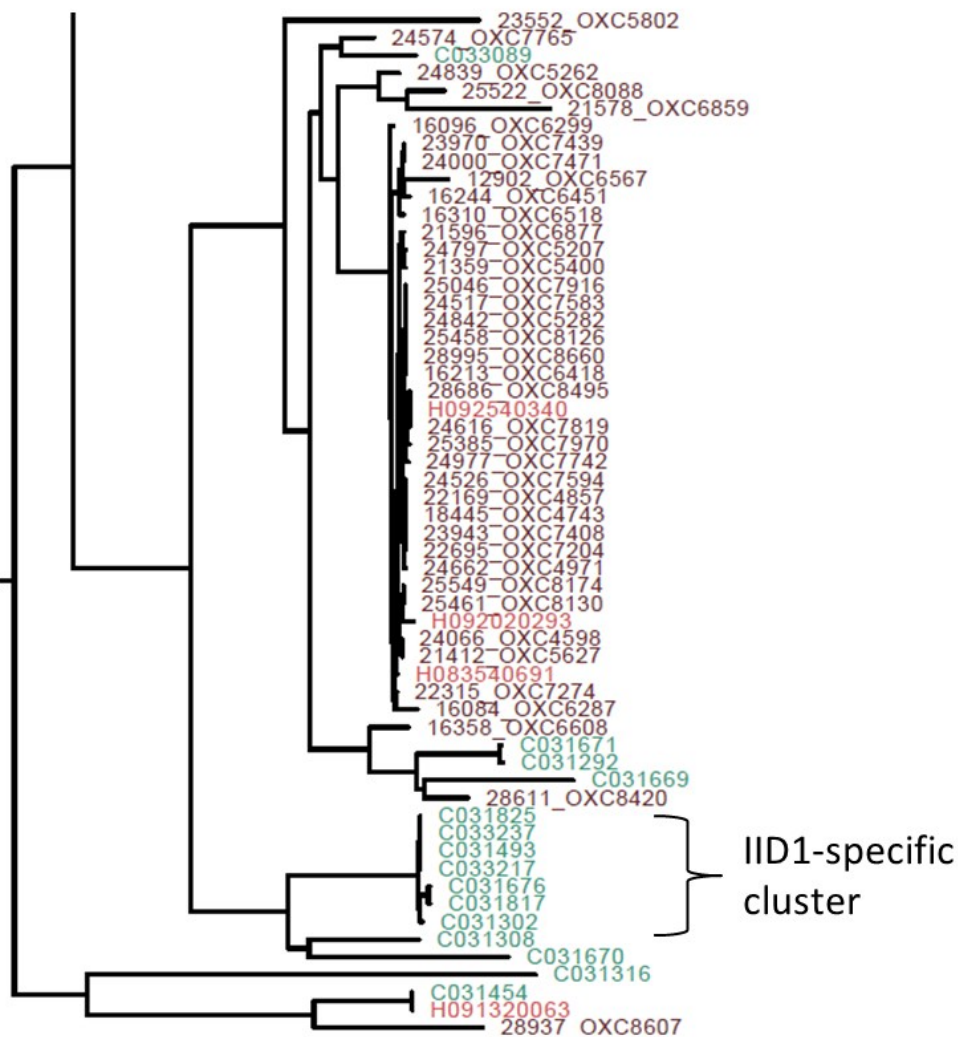


3.5.6 *C. coli* genomes

The 46 *C. coli* genomes generated during this project have been sent to Arnoud van Vliet (Institute of Food Research, Norwich) for further analysis as part of another FSA-funded project (FS101087) to investigate the molecular epidemiology of *C. coli* by next generation sequencing. Previous studies using WGS analysis (51) have demonstrated that *C. coli* can be sub-divided into three distinct clades. Most *C. coli* clinical isolates fall within Clade 3, which is associated with agricultural sources (51). Using SNP phylogeny, the *C. coli* isolates from IID1 and IID2 all cluster within Clade 1 (Arnoud van Vliet, personal communication).

Figure 6. IID-specific cluster within the broader *C. jejuni* SNP phylogeny. An IID1-specific cluster within one region of the phylogenetic tree shown in Appendix 7 is highlighted.

Key: IID1 isolates blue-green; IID2 isolates red; database isolates brown.



3.5.7 Other genomes

In addition to the *C. jejuni* and *C. coli* genomes, WGS data were also obtained for isolates from five *C. upsaliensis*, one *C. fetus*, one *Arcobacter butzleri* and one *Arcobacter* spp.

3.5.8 PacBio sequencing

3.5.8.1 Choice of strains for PacBio sequencing

We selected *C. jejuni* and *C. coli* strains for PacBio in order to obtain high quality genome sequences from major MLST clonal complexes or regions of the SNP phylogeny trees that are not currently represented by complete genomes in the public database. We also selected some from samples where clean genome sequence data had been difficult to obtain. In some cases, this was because the original samples contained more than one strain of *C. jejuni*. For four of the samples (two from each of IID1 and IID2), the PacBio sequence data was therefore derived from an isolate not represented amongst the Illumina dataset, adding to the overall number of isolates sequenced.

3.5.8.2 Preparation of samples for PacBio sequencing

Whilst considered far superior in terms of sequence read length and the possibility for obtaining either complete or near complete genomes (i.e. no gaps), the sample requirements for PacBio sequencing are far more stringent and we found this much more difficult to achieve for our *Campylobacteri* isolates. Ideally, a total of 10 µg DNA is required. In fact, we could only achieve this for six isolates. In addition, the DNA must be of very good quality. In particular, given the longer sequence read lengths that can be achieved, the DNA must be maintained in large fragments rather than showing signs of degradation.

In all, we attempted to prepare DNA from 26 different *Campylobacter* strains for PacBio sequencing. For each of the 26 strains, it was necessary to repeat the DNA extraction procedure 10 – 20 times in order to obtain a sufficient quantity of DNA. Even then, we had failures at the QC, library preparation and sequencing stages. After various troubleshooting experiments, we were able to overcome most difficulties by the addition of an extra wash step using the Magbead clean-up protocol (see section 3.4.5). Ultimately, we obtained PacBio sequence data from 14 isolates of *C. jejuni* and 3 isolates of *C. coli* (Table 3).

3.5.8.3 Analysis of PacBio sequence data.

The PacBio sequence data obtained is summarised in Table 3. The presence of a single large contig (> 1.6 Mb) matching *C. jejuni* or *C. coli* in the database indicates that the chromosome has assembled as a single contiguous DNA sequence for these strains (ten *C. jejuni* and one *C. coli*). Putative plasmids were identified in seven of the *C. jejuni* isolates (Table 3). Three

of these isolates were obtained from samples with mixed cultures (ie. more than one *C. jejuni*) and therefore represent different strains from the Illumina sequence data obtained from the same sample. These have been identified with the letter “b” at the end of the sample number and appear in Appendices 5 and 6 labelled as C031894b.pbio, H083540685b.pbio and C031457b.pbio.

3.5.8.4 Defining the core genome of *C. jejuni* using PacBio data

We carried out analyses using the 14 *C. jejuni* PacBio assemblies generated during this project and the three complete genomes available in the wider database (strains NCTC11168, RM1221 and 81-176). Relationships between the 17 strains, based on SNP phylogeny, are shown in Figure 7. The presence of a cluster (highlighted) for which no previous representatives had been completely sequenced demonstrates that we have contributed to an improved dataset of completed reference genomes. Based on these genomes, we defined a core genome for *C. jejuni*. Figure 8 shows a gene accumulation plot, demonstrating the reduction in the core genome as successive isolate genomes are included in the analysis. For these 17 isolate, a set of 1261 core genes (shared by all 17 isolates) was identified.

Figure 7 demonstrates that the addition of our PacBio genome sequence data considerably broadens the coverage of the *C. jejuni* population for which good quality reference genomes are available. Figure 8 demonstrates the reducing core genome and accumulating pan-genome as successive genomes are added into an analysis. The numbers of additional unique genes as each new genome is added (ie. present in just one *C. jejuni* genome) are low after the first few genomes, indicating that most of the variations between strains are due to the presence or absence of a restricted set of *Campylobacter* accessory genes, rather than constant influx of genes from outside the genus.

3.5.8.5 Variations in the carriage of virulence genes

We analysed the IID datasets, alongside 928 *C. jejuni* isolate genomes present in the pubmlst.org database, and previously sequenced reference genomes, to determine the prevalence of genes implicated in virulence. The results are shown in Appendix 8 and further demonstrate that there were no clear variations between the IID1 and IID2 collections, and that the IID collections are broadly similar to the collection already in the database.

Figure 7. *C. jejuni* SNP phylogeny based on the 14 PacBio-sequenced genomes and three complete genomes (in red) from the database.

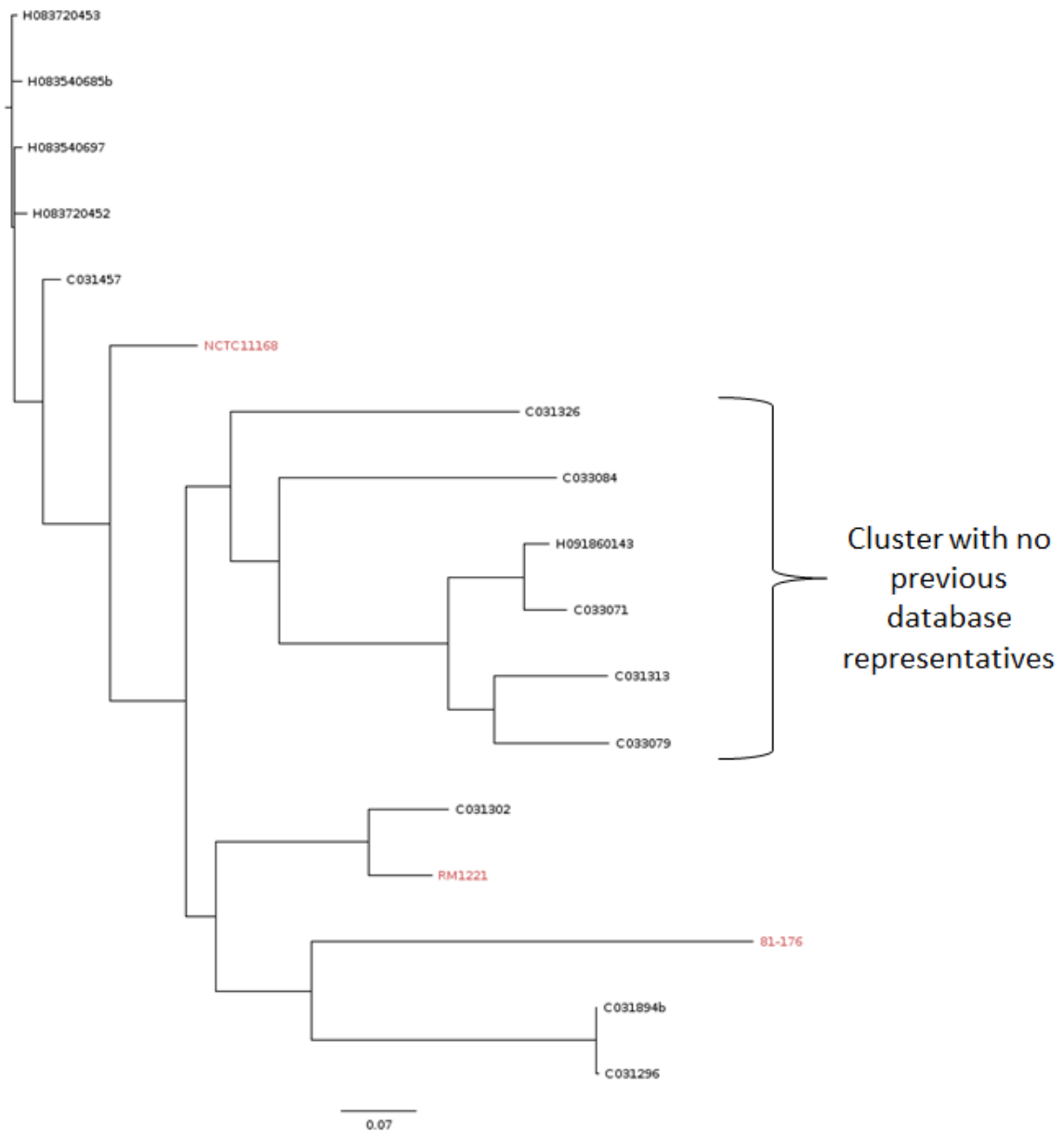
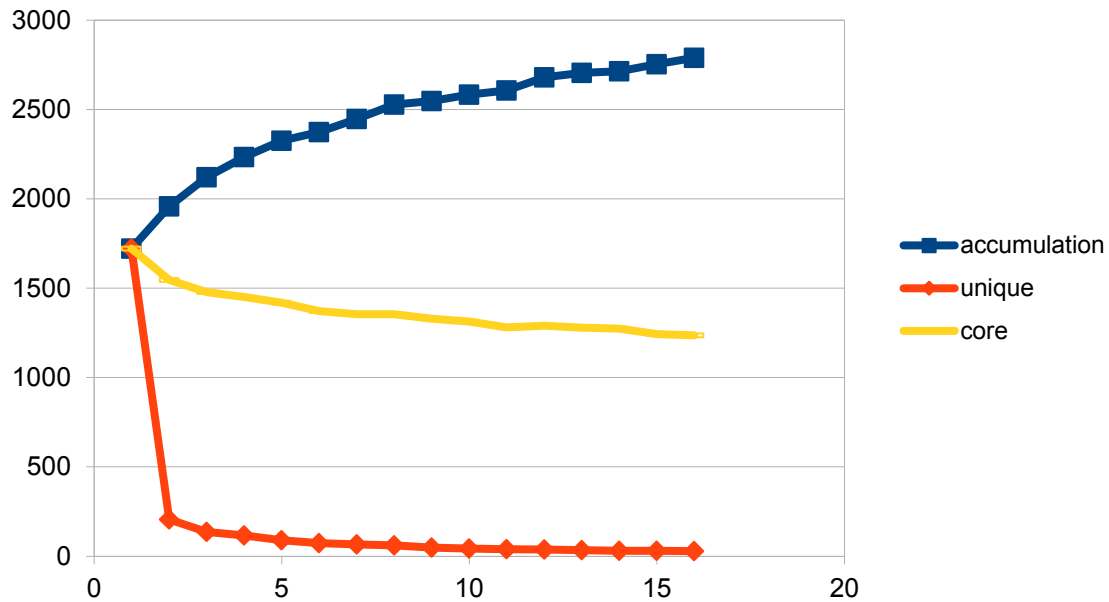


Figure 8. Gene accumulation plot based on the 14 PacBio-sequenced genomes and three complete genomes from the database. The y axis indicates numbers of open reading frames (ORFs) predicted as genes. The x axis plots the number of genomes used in the analysis. Unique refers to the number of ORFs found in just a single strain.



3.5.9 Database submissions

All data generated using Illumina and PacBio sequencing have been deposited to the European Nucleotide Archive (ENA) site (study number PRJEB7116).

Table 3. Summary of PacBio sequence data.

Strain	Study	ST	CC	Largest contig	Other contigs > 100 kb	Notes
<i>C. jejuni</i>						
C031296	IID1	257	257	1.73 Mb		Putative plasmid (55 kb)
C031302	IID1	324	354	1.75 Mb		
C033084	IID1	1361	-	1.76 Mb		
C031894b	IID1	257	257	1.71 Mb		pTet-like plasmid (43 kb)
H091860143	IID2	45	45	1.01 Mb	0.57 Mb	pTet-like plasmid (44 kb)
C031326	IID1	433	433	1.34 Mb	288 kb, 163 kb	163 kb contig with pTet-like sequences
H083720453	IID2	48	48	1.00 Mb	359 kb, 310 kb	
H083720452	IID2	48	48	1.67 Mb		
C033071	IID1	4024	45	1.67 Mb		Putative plasmid (51 kb)
H083540685b	IID2	48	48	1.75 Mb		Putative plasmid (12 kb)
C033079	IID1	45	45	1.63 Mb		pVir-like plasmid
H083540697	IID2	48	48	1.66 Mb		
C031457b	IID1	21	21	1.66 Mb		Putative plasmid (52 kb)
C031313	IID1	45	45	0.99 Mb	358 kb, 301 kb	
<i>C. coli</i>						
H08354_0696	IID2	830	828	1.00 Mb	405 kb, 278 kb	
H091860144	IID2	854	828	1.79 Mb		
H083540689	IID2	827	828	1.41 Mb	275 kb	

ST: (multilocus) sequence type

CC: clonal complex

3.6 Conclusions

The main conclusions from the study are as follows:

- (i) The overall breakdown according to MLST clonal complex for the IID1 and IID2 study isolates broadly reflects the distribution in the wider MLST database.
- (ii) There were no clear variations between the IID1 and IID2 collection based on their MLST profiles.
- (iii) Closer analysis using SNP-based comparisons of whole genome sequence data reveals some small clusters that are specific to either IID1 or IID2, and suggests that analysis based solely on MLST clonal complexes, rather than on sequence types, may be unreliable.
- (iv) Using PacBio sequencing, we have been able to add a further 17 high quality reference genomes to the database, eleven of which (ten *C. jejuni* and one *C. coli*) assembled as a single chromosome.

3.7. Recommendations for further work

- (i) These data provide an excellent baseline for monitoring shifts in the UK population of *Campylobacter* associated with gastrointestinal infections. By combining survey data of this nature with analyses of other isolate collections from non-human sources, it will be possible to identify changing trends and shifts in the relative importance of potential sources of transmission. This will require regular monitoring using WGS analysis. Contemporaneous monitoring of isolates associated with the primary source, poultry, would be required to confirm that any reductions in the incidence of campylobacteriosis are due to interventions aimed at the poultry industry.
- (ii) In our parallel study, we will be sequencing >1000 *C. jejuni* isolates from the environment (RCUK/FSA-funded *Campylobacter* research programme entitled “Sources, Seasonality, Transmission and Control: *Campylobacter* and human behaviour in a changing environment). A similar number of isolates from poultry and farm environments are being genome sequenced as part of the CamChain project (EMIDA-Era-Net scheme, EU). We will incorporate these data into a wider analysis of the two datasets, including Genome Wide Association Study (GWAS) and other statistical approaches to look for links between genome content and source, and infer the contribution of various sources to transmission to humans. These studies will be

completed in 2016. Any publications including the genomic data from the IID1 and IID2 studies will specifically acknowledge FS101072.

- (iii) We can employ similar approaches to look for correlations between gene content and strain metadata (for example, severity of infection), with the potential to identify specific strains associated with higher virulence, and determine any changes in their prevalence between IID1 and IID2.
- (iv) The additional high quality genomes provided to the community as a result of the PacBio sequencing in this project, can be exploited for improved assembly of short-read WGS data for epidemiological purposes.

3.8. Acknowledgements

We would like to thank the Food Standards Agency for funding this work (FSA Project FS101072). The following people have all contributed to this project: Fred Dove, Gosia Wnek, Carmen Martinez-Rodriguez (bacterial culture, DNA extractions); Margaret Hughes, Pia Koldkjaer (library preparation, QC).

3.9 References

1. Tam CC, Rodrigues LC, Viviani L, Dodds JP, Evans MR, Hunter PR, et al. Longitudinal study of infectious intestinal disease in the UK (IID2 study): incidence in the community and presenting to general practice. *Gut*. 2012;61(1):69-77.
2. Humphrey T, O'Brien S, Madsen M. Campylobacters as zoonotic pathogens: a food production perspective. *IntJFood Microbiol*. 2007;117(3):237-57.
3. Nachamkin I. Microbiologic approaches for studying Campylobacter species in patients with Guillain-Barre syndrome. *J Infect Dis*. 1997;176 Suppl 2:S106-14.
4. Wheeler JG, Sethi D, Cowden JM, Wall PG, Rodrigues LC, Tompkins DS, et al. Study of infectious intestinal disease in England: rates in the community, presenting to general practice, and reported to national surveillance. The Infectious Intestinal Disease Study Executive. *Bmj*. 1999;318(7190):1046-50.
5. Wassenaar TM, Newell DG. Genotyping of Campylobacter spp. *Appl Environ Microbiol*. 2000;66(1):1-9.
6. Clark CG, Bryden L, Cuff WR, Johnson PL, Jamieson F, Ciebin B, et al. Use of the oxford multilocus sequence typing protocol and sequencing of the flagellin short variable region to characterize isolates from a large outbreak of waterborne Campylobacter sp. strains in Walkerton, Ontario, Canada. *J Clin Microbiol*. 2005;43(5):2080-91.

7. Nielsen EM, Engberg J, Fussing V, Petersen L, Brogren CH, On SL. Evaluation of phenotypic and genotypic methods for subtyping *Campylobacter jejuni* isolates from humans, poultry, and cattle. *J Clin Microbiol.* 2000;38(10):3800-10.
8. Bang DD, Nielsen EM, Knudsen K, Madsen M. A one-year study of campylobacter carriage by individual Danish broiler chickens as the basis for selection of *Campylobacter* spp. strains for a chicken infection model. *Epidemiol Infect.* 2003;130(2):323-33.
9. Dingle KE, Colles FM, Falush D, Maiden MC. Sequence typing and comparison of population biology of *Campylobacter coli* and *Campylobacter jejuni*. *J Clin Microbiol.* 2005;43(1):340-7.
10. Dingle KE, Colles FM, Wareing DR, Ure R, Fox AJ, Bolton FE, et al. Multilocus sequence typing system for *Campylobacter jejuni*. *J Clin Microbiol.* 2001;39(1):14-23.
11. Dingle KE, McCarthy ND, Cody AJ, Peto TE, Maiden MC. Extended sequence typing of *Campylobacter* spp., United Kingdom. *Emerg Infect Dis.* 2008;14(10):1620-2.
12. Miller WG, On SLW, Wang G, Fontanoz S, Lastovica AJ, Mandrell RE. Extended multilocus sequence typing system for *Campylobacter coli*, *C. lari*, *C. upsaliensis*, and *C. helveticus*. *J Clin Microbiol.* 2005;43(5):2315-29.
13. Nichols GL. Fly transmission of *Campylobacter*. *Emerg Infect Dis.* 2005;11(3):361-4.
14. Sails AD, Swaminathan B, Fields PI. Clonal complexes of *Campylobacter jejuni* identified by multilocus sequence typing correlate with strain associations identified by multilocus enzyme electrophoresis. *J Clin Microbiol.* 2003;41(9):4058-67.
15. Manning G, Dowson CG, Bagnall MC, Ahmed IH, West M, Newell DG. Multilocus sequence typing for comparison of veterinary and human isolates of *Campylobacter jejuni*. *Appl Environ Microbiol.* 2003;69(11):6370-9.
16. Colles FM, Jones K, Harding RM, Maiden MC. Genetic diversity of *Campylobacter jejuni* isolates from farm animals and the farm environment. *Appl Environ Microbiol.* 2003;69(12):7409-13.
17. McCarthy ND, Colles FM, Dingle KE, Bagnall MC, Manning G, Maiden MC, et al. Host-associated genetic import in *Campylobacter jejuni*. *Emerg Infect Dis.* 2007;13(2):267-72.
18. French N, Barrigas M, Brown P, Ribiero P, Williams N, Leatherbarrow H, et al. Spatial epidemiology and natural population structure of *Campylobacter jejuni* colonizing a farmland ecosystem. *Environ Microbiol.* 2005;7(8):1116-26.
19. Karenlampi R, Rautelin H, Schonberg-Norio D, Paulin L, Hanninen ML. 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. *Appl Environ Microbiol*. 2007;73(1):148-55.

20. Taboada EN, Mackinnon JM, Luebbert CC, Gannon VP, Nash JH, Rahn K. Comparative genomic assessment of Multi-Locus Sequence Typing: rapid accumulation of genomic heterogeneity among clonal isolates of *Campylobacter jejuni*. *BMCEvolBiol*. 2008;8:229.
21. Wilson DJ, Gabriel E, Leatherbarrow AJ, Cheesbrough J, Gee S, Bolton E, et al. Tracing the source of campylobacteriosis. *PLoS Genet*. 2008;4(9):e1000203.
22. Sheppard SK, Dallas JF, Strachan NJ, MacRae M, McCarthy ND, Wilson DJ, et al. *Campylobacter* genotyping to determine the source of human infection. *Clin Infect Dis*. 2009;48(8):1072-8.
23. Sopwith W, Birtles A, Matthews M, Fox A, Gee S, Painter M, et al. Identification of potential environmentally adapted *Campylobacter jejuni* strain, United Kingdom. *Emerg Infect Dis*. 2008;14(11):1769-73.
24. Bronowski C, James CE, Winstanley C. Role of environmental survival in transmission of *Campylobacter jejuni*. *FEMS Microbiol Lett*. 2014;356(1):8-19.
25. Strachan NJ, Gormley FJ, Rotariu O, Ogden ID, Miller G, Dunn GM, et al. Attribution of *Campylobacter* Infections in Northeast Scotland to Specific Sources by Use of Multilocus Sequence Typing. *J Infect Dis*. 2009;199(8):1205-8.
26. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, et al. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology*. 2012;158(Pt 4):1005-15.
27. Read DS, Woodcock DJ, Strachan NJ, Forbes KJ, Colles FM, Maiden MC, et al. Evidence for phenotypic plasticity among multihost *Campylobacter jejuni* and *C. coli* lineages, obtained using ribosomal multilocus sequence typing and Raman spectroscopy. *Appl Environ Microbiol*. 2013;79(3):965-73.
28. Parkhill J, Wren BW, Mungall K, Ketley JM, Churcher C, Basham D, et al. The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*. 2000;403(6770):665-8.
29. Fouts DE, Mongodin EF, Mandrell RE, Miller WG, Rasko DA, Ravel J, et al. Major structural differences and novel potential virulence mechanisms from the genomes of multiple *campylobacter* species. *PLoS Biol*. 2005;3(1):e15.

30. Hepworth PJ, Ashelford KE, Hinds J, Gould KA, Witney AA, Williams NJ, et al. Genomic variations define divergence of water/wildlife-associated *Campylobacter jejuni* niche specialists from common clonal complexes. *Environ Microbiol.* 2011;13(6):1549-60.
31. Hofreuter D, Tsai J, Watson RO, Novik V, Altman B, Benitez M, et al. Unique features of a highly pathogenic *Campylobacter jejuni* strain. *Infect Immun.* 2006;74(8):4694-707.
32. Sheppard SK, Didelot X, Méric G, Torralbo A, Jolley KA, Kelly DJ, et al. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. *Proc Natl Acad Sci U S A.* 2013;110(29):11923-7.
33. Cody AJ, McCarthy ND, Jansen van Rensburg M, Isinkaye T, Bentley SD, Parkhill J, et al. Real-time genomic epidemiological evaluation of human *campylobacter* isolates by use of whole-genome multilocus sequence typing. *J Clin Microbiol.* 2013;51(8):2526-34.
34. Dorrell N, Mangan JA, Laing KG, Hinds J, Linton D, Al Ghusein H, et al. Whole genome comparison of *Campylobacter jejuni* human isolates using a low-cost microarray reveals extensive genetic diversity. *Genome Res.* 2001;11(10):1706-15.
35. Pearson BM, Pin C, Wright J, l'Anson K, Humphrey T, Wells JM. Comparative genome analysis of *Campylobacter jejuni* using whole genome DNA microarrays. *FEBS Lett.* 2003;554(1-2):224-30.
36. Leonard EE, Takata T, Blaser MJ, Falkow S, Tompkins LS, Gaynor EC. Use of an open-reading frame-specific *Campylobacter jejuni* DNA microarray as a new genotyping tool for studying epidemiologically related isolates. *J Infect Dis.* 2003;187(4):691-4.
37. Parker CT, Quinones B, Miller WG, Horn ST, Mandrell RE. Comparative genomic analysis of *Campylobacter jejuni* strains reveals diversity due to genomic elements similar to those present in *C. jejuni* strain RM1221. *J Clin Microbiol.* 2006;44:4125-35.
38. Colles FM, Maiden MC. *Campylobacter* sequence typing databases: applications and future prospects. *Microbiology.* 2012;158(Pt 11):2695-709.
39. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology.* 2012;19(5):455-77.
40. Frith MC, Hamada M, Horton P. Parameters for accurate genome alignment. *BMC bioinformatics.* 2010;11:80.
41. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403-10.

42. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30(14):2068-9.
43. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*. 2004;5:113.
44. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution*. 2009;26(7):1641-50.
45. Galardini M, Biondi EG, Bazzicalupo M, Mengoni A. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. *Source code for biology and medicine*. 2011;6:11.
46. Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J Bacteriol*. 2004;186(5):1518-30.
47. Sahl JW, Caporaso JG, Rasko DA, Keim P. The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. *PeerJ*. 2014;2:e332.
48. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC bioinformatics*. 2012;13:238.
49. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, et al. A whole-genome assembly of *Drosophila*. *Science*. 2000;287(5461):2196-204.
50. Chen L, Xiong Z, Sun L, Yang J, Jin Q. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic acids research*. 2012;40(Database issue):D641-5.
51. Sheppard SK, Didelot X, Jolley KA, Darling AE, Pascoe B, Méric G, et al. Progressive genome-wide introgression in agricultural *Campylobacter coli*. *Molecular ecology*. 2013;22(4):1051-64.

3.10 Appendices

Appendix 1: A summary of the WGS data generated using the Illumina platform.

Appendix 2: eBURST figure based on the MLST profiles of the IID1 and IID2 isolates, showing distribution amongst the wider population in the MLST database. eBURST is a tool for visualising population structures and relationships between strains based on MLST. Each dot represents a different sequence type. Two dots separated by a single line differ in one MLST locus (single locus variants). The size of a dot reflects its relative abundance. STs are colour coded as indicated to show the distribution of the IID1 and IID2 isolates (FSA dataset). Some common STs (ST-21, ST-45 and ST-257) are labelled. The eBURST algorithm assigns primary and sub-group founders to infer patterns of evolutionary descent. In the case of MLST, an example sub-group founder would be ST-21 for the ST-21 clonal complex.

Appendix 3: eBURST representation of just the IID1 and IID2 isolates. STs are colour coded as indicated to show the distribution of the IID1 and IID2 isolates (FSA dataset). Some common STs are labelled.

Appendix 4: Phylogenetic tree of IID1 and IID2 isolates based on rMLST loci. IID1 isolates, blue; IID2 isolates, red.

Appendix 5: Phylogenetic tree of IID1 and IID2 isolates based on core gene SNP phylogeny. IID1 isolates, blue; IID2 isolates, red; PacBio sequenced strains, green; database strains, black.

Appendix 6: Regions of the SNP-phylogeny tree from Appendix 5, highlighted to show the locations of MLST clonal complexes. Four different regions of the tree are shown. These are successively: upper, upper middle, lower middle and lower.

Appendix 7: Phylogenetic tree of IID1 and IID2 isolates and 928 other human isolates based on core gene SNP phylogeny. IID1 isolates, blue; IID2 isolates, red.

Appendix 8: Table showing percentage carriage of virulence genes amongst the IID1, IID2 datasets, the PacBio-generated genomes, three reference genomes (strains NCTC11168, RM1221 and 81-176) and the 928 *C. jejuni* human isolate genomes present in the pubmlst.org database. The numbers refer to percentage prevalence based on >90% identity, with partial matches (>25%) indicated in parentheses.