FINAL REPORT

Generating tools for the molecular epidemiology of *Campylobacter coli* by next generation genome sequencing

Project code: FS101087

Funder: UK Food Standards Agency

Contractor: IFR Enterprises Ltd

Project Lead Contractor: Dr Arnoud van Vliet

Report prepared by:

Arnoud H.M. van Vliet[1], Bruce M. Pearson[1], Nicola J. Williams[2], Ben Pascoe[3], Guillaume Meric[3], Phil Ashton[4], Claire Jenkins[4], Samuel K. Sheppard[3], Lisa C. Crossman[5,6], John Wain[7]

1. Institute of Food Research, Norwich, UK; 2. University of Liverpool, UK; 3. University of Swansea, UK; 4. Public Health England, Colindale, UK; 5. SequenceAnalysis.co.uk, Norwich, UK; 6. The Genome Analysis Centre, Norwich, UK; 7. University of East Anglia, Norwich, UK.

6 October 2016

CONTENTS

1 SUMMARY

Background & Aims: The foodborne bacterial pathogen *Campylobacter* is the most common cause of bacterial infectious intestinal disease in the UK, with an estimated 281,000 infections annually in the UK, and between 2 and 20 million cases annually in the European Union. Although *C. jejuni* is the most commonly isolated *Campylobacter*, approximately 10-15% of these infections are caused by *Campylobacter coli*. Despite the large number of cases associated with *C. coli*, relatively little is known about its environmental reservoirs, transmission routes or mechanisms of pathogenicity. Important differences exist between *C. coli* and *C. jejuni*, such as different infection sources, different antibiotic resistance profiles and different genetics, and yet the evidence base for the detailed investigation of *C. coli* is underdeveloped compared to *C. jejuni*. It has been shown by multilocus sequence typing (MLST) and comparative genomics that *C. coli* is clearly distinct genetically from *C. jejuni*, and is divided into distinct lineages representing agricultural, and environmental isolates. What is lacking for *C. coli* is sufficient resolution to allow the development of typing tools for rapid diagnostic and epidemiological investigation. With the recent developments in nucleic acid sequencing technologies it is now feasible to sequence large numbers of isolates to identify diagnostic markers and use these for future epidemiological purposes.
Results: In this project we have investigated the genetic diversity of *C. coli* using genome sequencing of 507 *C. coli* isolates from diverse sources including: animal, food, environment and human clinical cases. After assembly, 497 *C. coli* genome sequences were annotated, analysed, compared and have been made publicly available. These analyses have shown that *C. coli* has a population structure comprising of three ancestral clades, with Clade 1 containing two introgressed sequence clusters (Clusters A (ST-828) and B (ST-1150)), largely associated with agricultural sources and clinical samples, while the ancestral clades (Clade 1-Cluster C, Clade 2 and Clade 3) were mostly associated with environmental sources. However, a subset of Clade 3 isolates was associated with human disease, suggesting a link between environment and *Campylobacter* transmission. Genome coverage simulations were used to assess the coverage threshold required for the use of genome sequencing as a rapid, accurate and cost-effective tool for *Campylobacter* epidemiology. Lineage-specific genes have been identified and tested using other publicly available datasets.
Conclusions: *C. coli* has a very distinct population structure, with separation of lineages. The genome sequences and analyses provided by this project will support future studies into the biological differences between these lineages and their significance in human disease, and development of epidemiological tools for *C. coli*.

2 EXECUTIVE SUMMARY

The bacterium *Campylobacter* is the most common cause of bacterial food poisoning in the United Kingdom, with between a quarter and half a million cases annually. Similar figures are seen in other member states of the European Union, and reducing *Campylobacter*-related disease is one of the major goals of the Food Standards Agency and other funding bodies. There are two *Campylobacter* species which are primarily responsible for infections: *Campylobacter jejuni*, which is most commonly found on poultry meat and also in farm animals, and *Campylobacter coli*, which is also found on poultry meat but is thought to have more diverse set of environmental reservoirs. Approximately 10-15% of cases of *Campylobacter* illness is caused by *C. coli*, but despite this significant contribution to human disease, there has been relatively little attention given to *C. coli*, with most research focusing on *C. jejuni*. While these two bacterial species are very closely related, they are sufficiently different in their biology and transmission to warrant specific attention to be given to *C. coli*, and a major aim of this project was to generate the data required to allow for future *C. coli*-specific investigations, as well as generating new insights in the epidemiology of *C. coli*-caused disease.

Up to recently, the most commonly used technology to compare and trace *Campylobacter* isolates has been multilocus sequence typing (MLST), where a fingerprint is generated from 7 small conserved parts of the bacterial chromosome. While powerful, this technology does not have the resolution required to work with *C. coli*, as most isolates cluster together in what is called a clonal complex, named ST-828 for *C. coli*. Recent work with genome sequencing has demonstrated that *C. coli* is divided into distinct lineages (Clades) representing agricultural and environmental isolates. However, that was still based on a relatively small sample, and needed to be tested in a larger study with independent isolates. Also, what was still lacking for *C. coli* is sufficient resolution to allow the development of typing tools for rapid diagnostic and epidemiological investigation.

DNA sequencing technology has changed very rapidly over the last twenty years, and sequencing of genomes has now become both feasible and cost-effective for the characterisation of large numbers of *C. coli* isolates, with the goal of identifying diagnostic markers and use these for future epidemiological purposes.

The overall aim of this project was to use genome sequencing technology to determine and analyse the genetic diversity of *Campylobacter coli*, and to compare this diversity with information about where it was obtained from (agriculture, environment, clinical samples or food). Specific objectives were:

- to determine the genome sequence of 500 UK *C. coli* isolates from three different collections with a wide variety of isolation sources.
- to define the genetic variation within the *C. coli* genome.
- to compare the newly generated *C. coli* genomic information with existing typing methods.
- to support development of a typing scheme specific for *C. coli* using the information obtained from the genomes.

A total of 507 *C. coli* isolates were used for genome sequencing, with 497 used in further analyses. Almost all isolates originated from three independent UK collections. The isolates were subdivided in four categories: 1) 239 isolates from agricultural sources (farms, animals, poultry); 140 isolates from environmental sources (water, soil, ducks); 59 isolates from

clinical sources (blood, stool); and 59 from food (chicken, lamb, other food). The genomes were assembled from the sequencing data, the genes identified and then analysed using a diverse set of tools looking at diversity, ancestry and phylogeny. From the analyses it was very clear that *C. coli* has a population structure which is very different from that of *C. jejuni*, and has three ancestral clades (major groups), with the largest group having three different sequence clusters, of which Cluster A (ST-828) and Cluster B (ST-1150) clusters have acquired sequences from *C. jejuni*. Clade 1-Clusters A and B consisted mostly of agricultural, food and clinical isolates, whereas Clade 1-Cluster C and Clades 2 and 3 consisted mostly of environmental isolates. These five groupings are very different, suggesting they do not exchange DNA readily, and hence may not share the same environmental/agricultural niches. Genes specific for each Clade/Cluster have been identified, and compared to other datasets with regard to their usability in the future development of epidemiological tools. Importantly, this study addressed the underrepresentation of the environmental Clades in the available *C. coli* genome sequences, thus strengthening downstream analyses. Finally, datasets from this study were used to predict the minimally required amount of sequence data to be able to determine the genome sequence of *Campylobacter*, thus assisting future studies.

Conclusions: *Campylobacter coli* has a very distinct population structure with five clearly defined groups, with two groups (Clade 1-Clusters A and B) primarily associated with agricultural sources and human disease, while the other three groups (Clade 1-Cluster C and Clades 2 and 3) were mostly associated with environmental sources. *C. coli* has a very distinct population structure, with separation of lineages. The genome sequences and analyses provided by this project will support future studies into the biological differences between these groups, and assist in determining their significance in human disease.

2.1 GLOSSARY

- **Accessory genome:** The complement of genes not present in all isolates.
- **Antimicrobial resistance (AMR):** this describes the resistance to antibiotics and other antimicrobial compounds, resulting in these antibiotics being unable to treat infections.
- **Allele**: this is a variant form of a gene, commonly assigned a unique number or code.
- **Assembly:** the process of taking a large number of short DNA sequences and putting them back together to create a representation of the genome.
- **Clonal complex (CC):** A group of sequence types (STs) whose members are linked to at least one other member by being identical for four of the seven **MLST** genes.
- **Clade:** A grouping that includes a common ancestor and all the descendants (living and extinct) of that ancestor.
- **Contig:** A contiguous length of genomic sequence originating from the genome assembly process.
- **Core genome:** The complement of genes present in all isolates.
- **European Nucleotide Archive (ENA):** a database for DNA and associated sequences hosted by the EMBL European Bioinformatics Institute (EMBL-EBI, Europe).
- **Genbank:** a database for DNA and associated sequences hosted by the National Center for Biotechnology Information (NCBI, USA).
- **Genome sequencing:** the determination of the DNA sequence of the chromosome or chromosomes of an organism. Genomes can be complete (one contiguous sequence) or incomplete (draft, often called Whole Genome Shotgun (WGS)).
- **IID1:** Infectious Intestinal Disease Study 1 (1993 – 1996)
- **IID2:** Infectious Intestinal Disease Study 2 (2008 – 2009)
- **Illumina:** a short-read sequencing technology commonly used, delivering a large number of short reads that can then be assembled into a genome sequence.
- **Isolate**: A *Campylobacter* culture isolated from a specimen by microbiological methods.
- **Multilocus sequence typing (MLST):** a commonly used, DNA sequence-based typing method
- **L50:** a statistic relating to the length of contiguous sequences
- **N50:** a statistic relating to the length of contiguous sequences
- **PacBio:** Pacific Biosciences sequencing platform, a long-read sequencing platform.
- **Pangenome:** the full complement of genes in a clade.
- **Pulsed-Field Gel Electrophoresis (PGFE):** a gel-based low-resolution technique to separate large fragments of DNA, commonly used in older typing methods.
- **PHE:** Public Health England
- **Single Nucleotide Polymorphism (SNP):** a variation in a single nucleotide that occurs at a specific position in the genome.
- **Sequence type (ST)**: For each of the 7 genes used for MLST gene, the different sequences are assigned a unique number. The combination of the 7 alleles defines the sequence type (ST), for example 2-1-54-3-4-1-5 = ST-1 for *Campylobacter*.

3 FULL REPORT

3.1 INTRODUCTION

The foodborne bacterial pathogen *Campylobacter* is the most common cause of bacterial gastroenteritis in the UK and other industrialised countries, with an estimated 500,000 infections annually in the UK (1-3). The most commonly isolated *Campylobacter* species associated with intestinal infectious disease is *Campylobacter jejuni*, which is responsible for >50% of the human cases, whilst 10-15% of the cases are attributed to *Campylobacter coli*. Although both *Campylobacter* species are found in poultry, the major risk factor for *C. jejuni* is the consumption of chicken, there are specific risk factors for *C. coli* infection, including the consumption of game and tripe, and recreational swimming (4).

Sub-typing for *Campylobacter* in public health reference laboratories may still be based on phage typing, serotyping and pulsed field gel electrophoresis (PFGE), but these have not been used extensively over recent years because they do not, in general, provide useful information for outbreak investigation (1). This may be because of the epidemiology of *Campylobacter* (outbreaks may involve multiple sources), and/or because the population structure (extensive exchange of genetic material) makes interpretation of current typing data impossible. Multi Locus Sequencing Typing (MSLT) has been successfully established for both *C. jejuni* and *C. coli* (5, 6) and can be considered the gold standard, but although a highly sensitive tool, this typing method still gives insufficient information and resolution to distinguish between clinical, environmental and animal isolates of *Campylobacter*.

The rapid developments and reduction in cost in high-throughput next generation sequencing (NGS) now allow for the development of genomic-based analytical and epidemiological tools (7-10), as well as prediction of antimicrobial susceptibility profiles (11). Setting up genome sequencing-based applications for epidemiology requires the analysis of a large number of *Campylobacter* genomes to establish associations with disease, with host and with transmission routes, and the establishment of bioinformatic resources based on a thorough understanding of diversity, microbiology, epidemiology and evolution of Campylobacter. The overarching aim of this project is to contribute to the development of such tools for genome sequence-based molecular epidemiology.

***Campylobacter coli* genomics:** Source attribution for *C. jejuni* has been shown to be possible (12, 13), but understanding transmission routes requires epidemiological studies and an understanding of the biology of genome evolution. At the start of the project, there was no complete *C. coli* genome sequence available, and only 64 draft *C. coli* genome sequences were available in the Genbank/EMBL databases (14-17). These genome sequences did not allow for complete assessment of genome variability or evolution, and lack the required epidemiological data for downstream analysis. This means that, at the population level, some of the differences between *C. jejuni* and *C. coli* remained unclear. What was known is that *C. coli* has a wide range of infective sources, is associated with pigs, private drinking water supplies and foreign travel (4, 18). This all has direct relevance to food safety and it is scientifically unacceptable to assume that data from *C. jejuni* can be used to control infection caused by both species (6, 13, 19-21).

Population structure of *C. coli*: A study by Sheppard et al (16) showed that unlike *C. jejuni*, *C. coli* can be subdivided within clearly separated lineages, described as Clades 1, 2 and 3 based on gene content and sequence homology, accounting for introgression between species (Figure 1).

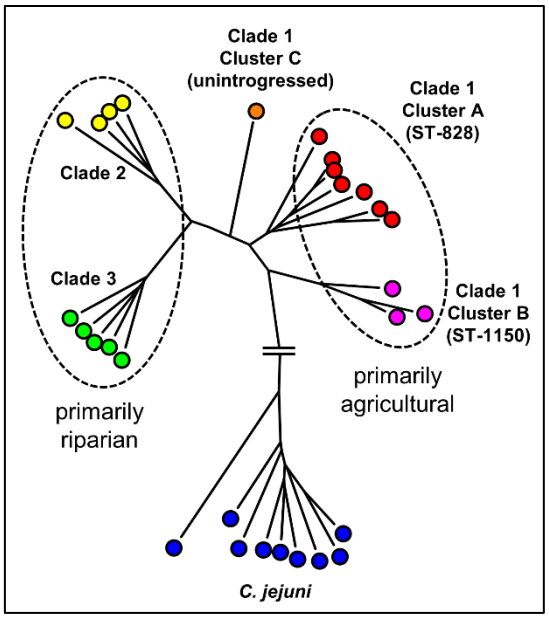


Figure 1. Schematic representation of population structure of *C. coli* and *C. jejuni* based on genome sequences, adapted from (16) and (22).

Clade 1 contained two sublineages, one consisting of the dominant MLST clonal complex ST-828, and one consisting of clonal complex ST-1150. This study was however based on a relatively small number of genome sequences (21 genomes). Analysis of genome sequences available via the Campylobacter PubMLST webpage (http://pubmlst.org/Campylobacter) resulted in a large increase of clonal complex ST-828 genome sequences, but very few Clade 2 and Clade 3 genome sequences, thus potentially representing a biased population due to the sampling strategy. As the Campylobacter PubMLST collection is heavily based on the Oxfordshire Campylobacter survey (7), this means it will primarily have clinical isolates and is unlikely to represent the true levels of genetic diversity in *C. coli*. In this project we have sequenced *C. coli* isolates from the UK, with emphasis on a wide variety of sources, including agriculture, water, food, wild birds and clinical cases.

Scientific approach: Strain selection is key to the success of this project. The consortium has included three UK-based groups with large, well defined *C. coli* strain collections (Public Health England, University of Liverpool and University of Swansea), significant expertise in *Campylobacter* genomics and epidemiology (Swansea, IFR, PHE, Liverpool) and genome sequencing and tools development (TGAC, UEA). The genome sequence of at least 500 *C. coli* isolates would be determined, and used for comparative genomics and phylogenetic analyses. The genome sequences would be made publicly available to support ongoing initiatives to combat *Campylobacter* foodborne illness.

Advancement of knowledge for Evidence Requirement: Three major areas of progress were anticipated from this project: (1) A better understanding of the contribution of *Campylobacter coli* in human infections in the UK; (2) development of a set of epidemiological tools based on high-throughput genome sequencing and whole genome sequences, which allow for tracking transmission through the food chain and veterinary sources; (3) Improved understanding of the genetic basis of the *C. coli* clade structure and the adaptive importance of inter-species recombination between *C. coli* and *C. jejuni* in the emergence of clinically important lineages.

Alignment with FSA policy needs: This project is aligned with several of the major targets identified in the 2009 "Joint Funders UK Research and Innovation Strategy for Campylobacter in the food chain, 2010-2015":

- Development of novel detection and diagnostic tools, and resources
- A strain bank to assist in understanding the genetic diversity of the bacterium

Furthermore, the research carried out here will also be informative for use of next generation sequencing for epidemiology and evolution of other foodborne bacterial pathogens, hence making a general contribution to FSA targets in microbial food safety.

3.2 AIMS OF THIS PROJECT

1. to determine the genome sequence of 500 *C. coli* isolates from different sources
2. to define the genetic variation within the core and accessory regions of the genome
3. to compare the newly generated *C. coli* genomic information with existing typing methods
4. to support development of a typing scheme specific for *C. coli* using the information obtained from the genomes

3.3 EXPECTED OUTCOMES OF THIS STUDY

Studying *C. coli* whole genome sequences will allow us to better understand specific differences in transmission and will provide a platform for the study of the genetics of transmission and pathogenicity by defining the phenotype of closely related strains. Downstream studies on phenotype-genotype associations can then be initiated. Furthermore, these studies will inform on the nature of host association and potential targets for antimicrobial or biological intervention in disease progression or transmission.

Envisaged outcomes were:

- Collection of 500 genome sequences of *Campylobacter coli*, isolated from agricultural, animal, food, environmental and clinical sources.
- Curated database of genome sequences for epidemiological and evolutionary analyses.
- Final report with recommendations on usage of whole genome sequencing for use in molecular epidemiology of *C. coli.*

3.4. EXPERIMENTAL PROCEDURES

3.4.1 Study design

C. coli isolates with corresponding clinical, source or other metadata, including typing data, were selected from three well-defined collections: (i) clinical and environmental isolates from the Public Health England, (ii) agricultural, environmental and food isolates from the Universities of Liverpool and Swansea. Isolates were cultivated and stored at IFR, and genomic DNA was isolated and used for high-throughput sequencing on the Illumina HiSeq platform at TGAC. Raw sequence data was passed through the TGAC quality control pipelines, assembled and annotated at IFR, and analysed by the participants. A complete, finished genome sequence was obtained for representatives of each separate clade, and used for comparative genomics and molecular epidemiology analyses (see specific sections).

3.4.2 Bacterial strains and growth conditions

A total of 507 *C. coli* isolates were included in this study, and were retrieved from strain collections from University of Liverpool (N=237), Public Health England (N=183), University of Swansea (N=58), University of Nottingham (N=24), University of Bristol (N=4) and the Institute of Food Research (N=1). Isolates were selected to represent the wide variety of sources for *C. coli*, and were primarily from the United Kingdom. A more detailed breakdown of isolates, sources and other information is provided in the Results section. Isolates were shipped on dry ice or in Transwab® Amies Charcoal Transport tubes (Medical Wire & Equipment Co. Ltd, Corsham, Wiltshire, UK), and subsequently cultivated on Brucella agar plates (Becton, Dickinson and Company) at 37ºC in microaerobic conditions (85% N_2, 10% CO_2, 5% O_2) in a MACS-MG-100 controlled atmosphere cabinet (Don Whitley Scientific), and subsequently stored at -80ºC using -80C Protect® bead stocks (Technical Service Consultants Ltd, Microbiology House, Fir Street, Heywood, Lancashire, UK). After the first 101 isolates, all samples were tested using *C. coli* and *C. jejuni* primer sets, as older discrimination methods (PCR and hippicurate tests) are not 100% reliable and 10 isolates originally classified as *C. coli* turned out to be *C. jejuni*. The following primer pairs were obtained from (23), and were successfully used to confirm that all the remaining isolates were *C. coli*.

Species	Forward primer	Reverse primer	Product size
C. coli	GGTATGATTTCTACAAGCGAG	ATAAAAGACTATCGTCGCGTG	502 bp
C. jejuni	GAAGAGGGTTTGGGTGGTG	AGCTAGCTTCGCATAATAACTTG	735 bp

3.4.3 Isolation of genomic DNA

Isolates were grown overnight at 37ºC in microaerobic conditions in Brucella broth (Becton, Dickinson and Company), and cells were harvested by centrifugation at 3000×*g* for 10 min. Genomic DNA was isolated by a standard method (24). Briefly, cells were resuspended in Tris-EDTA buffer, lysed by addition of SDS, followed by incubation with proteinase K at 65°C for 10 minutes. Subsequently, 5M NaCl and 10% CTAB (hexadecyltrimethyl ammonium bromide) in 0.7 M NaCl were added, followed by incubation at 65°C for 10 minutes. DNA was extracted using chloroform/isoamyl alcohol (24:1) and precipitated using isopropanol, washed with 70% ethanol and resuspended in TE with DNase-free RNase and incubated at 37°C for 30 minutes. Genomic DNA was ethanol precipitated and resuspended in TE buffer, and the concentration was determined by Nanodrop.

3.4.4 High-throughput DNA sequencing using Illumina technology

Genomic DNA samples were sent to The Genome Analysis Centre (TGAC), Norwich, UK, for QC and library construction. Libraries for samples Cc42yr and LDI4896 to LDI9961 were generated using the Illumina TruSeq kit, according to the manufacturers instructions (Illumina UK, Cambridge). Libraries for samples LDI12891 to LDI14948 were generated using the Illumina NexteraXT kit, according to the manufacturers instructions (Illumina UK, Cambridge). The change in protocol was due to changes in sequencing technology, and the NexteraXT kit has the advantage of requiring significantly less DNA when compared to the TruSeq kit. The average library size was between 400-600 bp. The libraries were used for high-throughput DNA sequencing using the Illumina HiSeq 2500 in rapid run mode with 100 nt paired end reads (samples LDI4896 to LDI13635) or on the Illumina MiSeq with 100 nt paired end reads (sample Cc42yr) and 250 nt reads (samples LDI14925 to LDI14948). Reads were QC'ed, trimmed and adaptors removed using standard protocols, and released as compressed FASTQ files.

3.4.5 High-throughput DNA sequencing using PacBio technology

After the first run of 20 genomes, 5 isolates were selected to generate complete genome sequences using the PacBio long-read sequencing technology. Genomic DNA purified as described in section 3.4.3 was used for PacBio sequencing at TGAC using 2 SMRT cells per sample, resulting in ~50,000 to 100,000 reads per sample.

3.4.6 Assembly of draft and complete genome sequences

Genome sequences were assembled using SPAdes 3.5.0 (25). Initially, Velvet (26) was used, but Velvet does not allow hybrid assembly of Illumina and PacBio data, and also showed lower assembly efficiency when compared with SPAdes. This is consistent with publications benchmarking assembly software (27), hence all sequences were re-assembled using SPAdes, using the following settings: [1] k-values of 33, 55, 77, 99 (for 100 nt paired end reads) and 33, 55, 77, 99, 127 (for 250 nt paired end reads); [2] --careful and --mismatch-correction switches for high quality assemblies; [3] --cov-cutoff 10 to avoid inclusion of low coverage contigs. For hybrid assembly of Illumina and PacBio sequencing data, the --pacbio switch was included. Assembly statistics were obtained by analysing the resulting assemblies using Quast v. 2.0 (28).

3.4.7 Genome coverage simulation

Coverage (read depth or depth) is the average number of reads representing a given nucleotide in the reconstructed sequence. It can be calculated from the length of the original genome (G), the number of reads (N), and the average read length (L) as (N*L)/G. We have used Subsampler (https://github.com/homonecloco/subsampler) to randomly sample a proportion of reads of 4 different *C. coli* sequencing samples (LDI6736, LDI6746, LDI6766, LDI6786) which represent samples with a low number of reads (LDI6736) to samples with higher than average number of reads. Overall these samples represented a range of 24.5× - 159.8× fold sequence coverage. We sampled 10%, 25%, 50% and 75% of the reads, calculated the coverage compared to the 100% reads dataset, and plotted the sequence coverage vs the percentage of the genome assembled.

3.4.8 Annotation of genome sequences

Contigs obtained from SPAdes assembly were reordered using Mauve v 2.4.0 (29, 30), using the complete genomes of *C. coli* 15-537360 (Clade 1-Cluster A), *C. coli* C8C3 (Clade 2), *C. coli* C9A1 (Clade 3), *C. coli* N11D1 (Clade 1-Cluster C), *C. coli* RM4661 (Clade 1-Cluster B) and *C. jejuni* 81-176 (jejuni) as reference. Reordered genomes were annotated using Prokka v.

1.11b (31). Prokka combines the following programs: [1] Prodigal for identification of open reading frames (ORFs), [2] BLAST+ and HMMER for homology searches, [3] Aragorn for identification of transfer RNA genes, [4] Barrnap for identification of ribosomal RNA genes, [5] MINCED for identification of CRISPR repeats, and [6] tbl2asn for generation of Genbank-compliant GBK and GFF3 files. Prokka does not annotate incomplete ORFs on the edge of contigs.

3.4.9 Phylogenetic analyses

Core genome single nucleotide polymorphisms (SNPs) were identified using the parSNP program v. 1.2 from the Harvest suite (32) with the "-a 13 -c -x" switches, and used for the generation of phylogenetic trees using the default settings of parSNP. Whole genome-based phylogeny was determined using Feature Frequency Profiling (33), with the FFP version 3.19 suite of programs (http://sourceforge.net/projects/ffp-phylogeny/) (34), utilising the FFPry program. These FFP programs generate a distance matrix, with phylogenetic trees being generated using the Neighbor Joining algorithm. Multilocus sequence typing (MLST) was performed as previously published for core genome, pan genome, ribosomal genes and the original 7-gene MLST schemes (35). Figtree (http://tree.bio.ed.ac.uk/software/figtree/) was used for annotation of phylogenetic trees. *C. coli* genomes were subdivided in lineages based on the Clade designation (Clade 1-Cluster A (ST-828), Clade 1-Cluster B (ST-1150), Clade 1-Cluster C (unintrogressed), Clade 2 and Clade 3) described previously (16, 22), based on subdivision of the genomes used in those studies.

3.4.10 Comparative genomics analyses

Genomes were first searched for the presence/absence of a panel of known variable genes in *C. jejuni* and *C. coli*. Gene sequences were subdivided in 120 nt fragments and for *in silico* genotyping using the Microbial In Silico Typing (MIST) software package (36) and the NCBI Blast+ v2.28 executables (see Supplementary Table S4-S6 for details).

Pan/Core/Accessory genomes and identification of clade-specific genes: The pangenome of the 507 genomes was determined using Roary v3.5.8 (37). Roary uses CD-hit (38) to speed up searching and allows for multiple comparisons. Outputs include a tab-delimited file with annotations, as well as other files for visualisation of accessory and core genomes. Due to the sequence divergence between the *C. coli* lineages, Roary was used using the -s switch which groups paralogs and reduced separation of allelic differences, and was used with a cut-off of 80% identity (-i 80 setting) with clustering of paralogs (-s setting). Due to the high levels of genetic variability between the clades and sequence clusters, we used more relaxed criteria to search for clade/cluster-specific genes. The complete genome sequences determined in this study (containing one Clade 1-Cluster A (ST-828), one unintrogressed Clade 1-Cluster C genome, two Clade 2 genomes and one Clade 3 genome) were used as reference genomes.

3.4.11 Identification of core and accessory genomes of *C. coli*

C. coli genomes determined in this study were combined with genomes present in the Campylobacter PubMLST database, resulting in 692 genomes from Clade 1-Cluster A (ST-828), 12 from Clade 1-Cluster B (ST-1150), 46 from Clade 2, 60 from Clade 3 and 35 from unintrogressed Clade 1-Cluster C (6, 16). We completed this dataset with 362 *C. jejuni* isolates from various published (39, 40) and unpublished works to capture as much as possible of the whole diversity of the two species. A phylogenetic tree was constructed for these 1207 genomes, based on a published core genome for both species (40). To characterise core and accessory genome variation, a previously published list of all genes present in 7 reference genomes of *C. jejuni* and *C. coli* was used (10). Briefly, the list of all coding sequences (CDS) found in all 7 reference genomes was filtered using BLAST to keep only the

CDS uniquely present in the dataset, removing all alleles and duplicates for each gene. The subsequent list contained 2335 genes, which are contained reference sequences for all genes present in 7 reference strains, which provides a good starting point for core and accessory genome analysis, as we have published in 2014 (10). All genes from this list were used to scan all 853 *C. coli* genomes of this project, comprising genomes sequenced in this study as well as genomes published in other studies (10, 16).

3.4.12 Prediction of antimicrobial resistance

Prediction of antimicrobial resistance in *C. coli* genome sequences (11, 41) was done using several independent approaches: [1] Presence/absence analysis using known *Campylobacter* antimicrobial resistance genes for tetracycline, aminoglycosides, chloramphenicol; [2] analysis of the 23S rRNA region containing the residues at position 2074 and 2075, where mutations such as $A_{2074}G$ and $A_{2075}G$ are known to cause resistance to macrolides/lincosamides/ketolides (MLKs); [3] analysis of the predicted GyrA amino acid sequence, where $T_{86}I$ is associated with quinolone resistance; [4] Analysis of the FASTQ reads using SRST2 (42) and the curated ARGannot and ResFinder databases provided at https://github.com/katholt/srst2.

3.4.13 Data availability

The FASTQ reads, isolate information and genome assemblies have been uploaded to the European Nucleotide Archive (ENA) at the EMBL-EBI (http://www.ebi.ac.uk/ena), as project PRJEB11972. The genome sequences have been uploaded into the Campylobacter pubMLST databases. Accession numbers for ENA (reads, assemblies) and Campylobacter pubMLST database are provided in Supplementary Table S3. Annotated genome sequences are available via http://www.ifr.ac.uk/Campylobacter/campylobactercoligenomics.html.

3.5 RESULTS

3.5.1 *Campylobacter coli* isolates

A total of 617 *C. coli* isolates were collected for this project between October 2013 and July 2015, and included contribution of *C. coli* isolates from two other UK universities (University of Nottingham and University of Bristol). The isolates were selected to represent a wide variety of isolation sources, and were all isolated in the United Kingdom. Of these, 512 were used for genome sequencing, with the decision on isolates sequenced in part based on availability during the timeframe of the project, and to include a significant proportion of isolates from non-agricultural/non-clinical backgrounds. Of the 512 isolates sequenced, 7 were mixed samples (either *C. jejuni* and *C. coli* or two *C. coli* isolates), of which 2 were restreaked to single colonies, purified and sequenced again. After the first two sequencing runs, 10 isolates were found to be *C. jejuni* rather than *C. coli*. This misidentification is not uncommon and subsequently PCR was used to distinguish *C. coli* and *C. jejuni* isolates prior to genome sequencing (see section 3.4.2), ensuring the remainder of isolates characterised were only *C. coli*. The final total of *C. coli* isolates was 497, with 10 *C. jejuni* isolates. Table 1 shows an overview of the *C. coli* isolates separated per contributor, including the source of isolation subcategory (agricultural, environmental, food and clinical), whereas Table 2 shows an overview of the *C. coli* isolates separated per isolation source of each subcategory. Supplementary Table S1 lists all information of the isolates.

Table 1. *C. coli* isolates for which the genome sequence has been determined in this study

Contributed by:	Agriculture	Environment	Clinical	Food	Total
Liverpool University	177	51	0	0	228
Public Health England	2	66	56	59	183
Swansea University	36	22	0	0	58
Nottingham University	21	1	2	0	24
Bristol University	3	0	0	0	3
IFR	0	0	1	0	1
Total	239	140	59	**59**	**497**

Table 2. Overview of *C. coli* isolates per isolation source subcategory

	Chicken	Farm environment	Dog	Pig
Agricultural (N=239)	53	178	1	7
	Duck	Environmental water	Soil	Other
Environmental (N=140)	23	81	35	2
		Human unspecified	Human stool	Blood
Clinical (N=59)		5	7	47
	Chicken	Lamb	Pork	Other food
Food (N=59)	39	6	3	11

Isolation source subcategories based on the Campylobacter pubMLST (BIGSdb) classification.

3.5.2. Genome sequencing: overview and statistics

A total of 482 *C. coli* DNA samples was sequenced using the Illumina HiSeq platform with 100 nt paired end reads, whereas 25 samples were sequenced using the Illumina MiSeq platform with 100 nt paired end reads (N=1) and 250 nt paired end reads (N=24). To establish the minimum sequence coverage required for genome assembly of *Campylobacter*, 4 different *C. coli* sequencing samples (LDI6736, LDI6746, LDI6766, LDI6786) representing a range of

24.5× - 159.8× fold sequence coverage. Using Subsampler, these FASTQ files were randomly sampled at 10%, 25%, 50% and 75% of the reads, assembled using Velvet and the coverage of the assembled genomes was compared to the 100% reads dataset, and plotted the sequence coverage vs the percentage of the genome assembled (Figure 2).

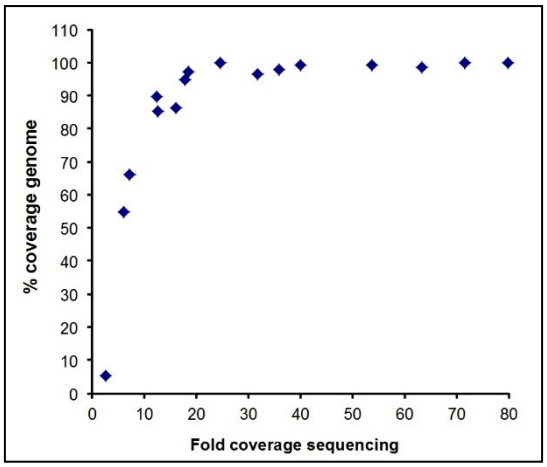


Figure 2. Sequence coverage vs percentage of the *Campylobacter* genome assembled, based on random subsampling of 4 *C. coli* genome sequencing samples.

When sequence coverage dropped below 10-fold (approx 85,000 paired end, 100 nt reads), less than 80% of the genome was assembled. Also, the level of fragmentation (estimated by the number of contigs generated) increased significantly, >10-fold (not shown). Hence the minimum coverage required for *Campylobacter* genome sequencing is 10-fold, but we recommend that sequence coverage is where possible kept >15-fold, preferably >20-fold. This was achieved with all samples sequenced (lowest was 24.5-fold coverage with >450 samples had >50-fold coverage). The improvements in Illumina sequencing technology that have occured during the project make it unlikely that such low sequence coverage will occur, and hence is not expected to cause future problems. Further assemblies were performed using SPAdes v. 3.5.0, as this assembler was able to work with genomes where Velvet failed to assemble the full genome. Assembly statistics for individual samples are provided in Supplementary Table S3 and are further discussed in Section 3.5.6.1.

Short-read sequencing technologies such as those offered by Illumina do not allow completion of genome sequences as the reads are to short to resolve repeated sequences and duplicated areas, such as the rRNA operons. As completed genomes are better suited for use as reference genomes, we used the long-read technology offered by PacBio sequencing to complete 5 genomes representing the different *C. coli* lineages (see section 3.5.4). The genomes of isolates 15-537360 (Clade 1-Cluster A, ST-828), N11D1 (Clade 1-Cluster C, unintrogressed), M5D4 and C8C3 (Clade 2), and C9A1 (Clade 3) were completed into a single contig for the genome, and separate plasmids where these were present.

3.5.3 Multilocus sequence typing of *C. coli* genome sequences

The MLST scheme from the Campylobacter PubMLST website was used to determine the 7-gene MLST-alleles for the 497 *C. coli* genomes (Table 3 for a summary, Supplementary Table S2 for full data). The majority of isolates (n=318) had a sequence type (ST) which is part of the ST-828 clonal complex, whereas 8 isolates are part of the ST-1150 clonal complex. There were 88 isolates with an ST which not part of a clonal complex, while the remainder (n=85) has a yet unclassified ST.

Table 3. Distribution of MLST sequence types and clonal complexes for *C. coli* isolates

Clonal Complex	Number of isolates	Sequence type	Number of isolates
ST-828	318	825	33
		827	187
		828	17
		Other sequence type	81
ST-1150	8	1487	8
No clonal complex	171	Other sequence type	88
		Unclassified	83

The distribution of MLST sequence types and clonal complexes shows the known dominance of the ST-828 clonal complex in *C. coli* collections (62.7%), but also that the sequences obtained contain a large proportion (34.1%) of sequence types not part of a clonal complex, when compared to publicly available genome sequences, where 106 out of 657 (16.1%) is not part of a clonal complex, and 546 out of 657 (83.1%) is part of the ST-828 clonal complex.

The gene-by-gene approach adopted in this project is based on the archiving and analysing of thousands of genomes using the web-based BIGSdb platform (43), also called the whole-genome MLST approach (9). This means that it is possible to configurate a link between the genome archive and the Campylobacter pubMLST database, servers (http://www.pubmlst.org/campylobacter/) and schemes. When MLST target genes can be detected, i.e. when their sequences are not located at the end of assembled contiguous sequences, an allelic type matching the allelic types stored in the pubMLST database will be given, as well as a clonal complex, if existing. In a dataset of 1,207 genomes, we were able to assign allelic types for *aspA*, *glnA*, *gltA*, *glyA*, *pgm*, *tkt* and *uncA*, and therefore sequence types (STs) for 1,113 genomes, or 92.2% of the dataset. A total of 94 genomes, mostly *C. coli* clades 2 and 3, could not be typed. This could be caused by the fragmented nature of genome assemblies, but more likely because isolates from these clades are still poorly characterised in public databases, and no allele designations or sequence types exist for these lineages yet.

The 739 isolates from *C. coli* clade 1 were grouped into three groups: ST-828 complex (Cluster A), ST-1150 complex (Cluster B), and unintrogressed isolates (Cluster C) (6, 16), which corresponded to a total of 151 distinct STs. The most common ST in Clade 1 was ST827 (ST-828 clonal complex), followed by ST825, ST828, ST1090 and ST872. We found 3 distinct STs in ST-1150 complex: the dominant ST1487, but also ST3667 and ST5160. Only 3 unintrogressed clade 1 isolates could not be typed (3/35, 8.5%) and the rest belonged to ST3311, ST1764, ST1765, ST1986 and other minor STs. Twenty-four out of 46 (52%) clade 2 isolates could not be typed, with the rest belonging in majority to ST3319, ST3318, ST1988 and other minor STs. Fifty out of 60 clade 3 isolates could not be typed (83%), with the rest being typed to ST3309, ST1576, ST1992, ST6133 and other minor STs. All MLST typing results are presented in Supplementary Table S2. All genome sequences from this project have been uploaded in the Campylobacter pubMLST database, allowing future re-analysis once allele designations have been assigned, and potential clonal complexes have been identified.

This result justifies the use of a more global, post-genomic approach to molecular typing, which is very well integrated in the whole-genome MLST approach implemented in BIGSdb (44). Using the approach explained in section 3.4.11, we could observe allelic variation in our dataset at all loci examined, with loci showing very limited allelic variation across the whole dataset (typically ribosomal proteins and conserved housekeeping genes) and loci showing very high variation. For instance, the *porA* gene, encoding for a major outer membrane protein (45), had very high allelic variability in the dataset, with a new allele of the gene harboured every 3 isolates examined, in average. More interestingly and because of the

population structure of the *Campylobacter* genus, many genes showed a significantly higher allelic variation in *C. coli* than in *C. jejuni*. Amongst the top genes were capsular polysaccharides genes *kpsS* and *kpsD*, as well as the cytolethal distending toxin subunit *cdtB* and formate dehydrogenase *fdhA*. These genes had about 20% more sequence divergence in *C. coli* than in *C. jejuni*, which could indicate a possible bottleneck of diversity in *C. jejuni* at these loci, which have been shown to be important for host colonisation (46-48). More generally, could be interesting targets to explain variations in natural selective pressures acting on these genes in two species sharing the same ecological niche.

3.5.4 Assignment of genome sequences to lineages and comparison with publicly available *C. coli* genome sequences

The genomic diversity within the 497 genome sequences of this study was determined by identification of single nucleotide polymorphisms (SNPs) followed by phylogenetic clustering, by core genome MLST (7) and by using whole genome comparisons using Feature Frequence Profiling (FFP) (33). The phylogenetic trees obtained by parSNP and FFP were comparable in clustering, and hence only the SNP-based trees are shown here. We first assigned the *C. coli* genomes to the lineages proposed previously (16, 22): the agricultural lineages Clade 1-Cluster A (ST-828), Clade 1-Cluster B (ST-1150); Clade 1-Cluster C (unintrogressed *C. coli*), and the environmental (riparian) lineages Clade 2 and Clade 3 (Figure 3). Assignment to these lineages was done by including the 21 *C. coli* and 8 *C. jejuni* genome sequences used originally (16).

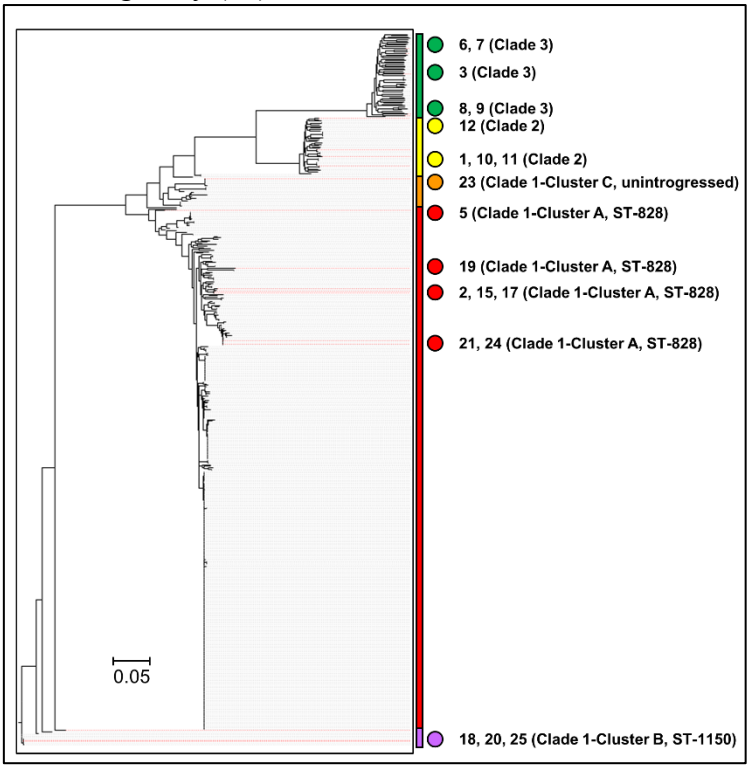


Figure 3. Phylogenetic tree comparing the 497 *C. coli* and 10 *C. jejuni* sequences with the 21 *C. coli* genome sequences from (16), for assignment to lineages Clade 1-Clusters A-C, and Clades 2 and 3. The color bar to the right shows the delineation of the lineages, the numbers are those from the reference sequences from (16).

The assignment to different lineages is consistent with the deliniation using core genome MLST (Figure 4).

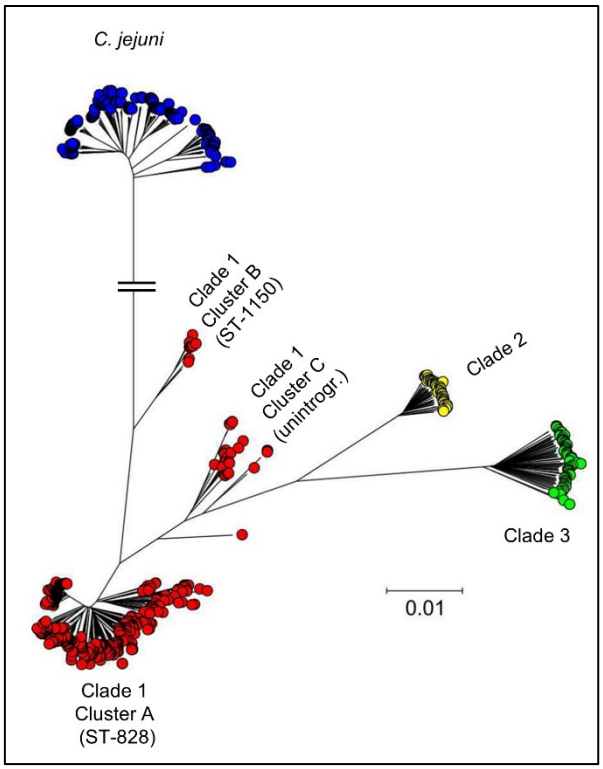


Figure 4. Phylogeny of *C. coli* and *C. jejuni* used in this project. A core genome alignment, based on a previously published list of 595 genes shared by *C. jejuni* and *C. coli* (40) was produced for 1207 genomes and used to create a phylogenetic tree using the neighbour joining algorithm. Blue circles represent 362 genomes of *C. jejuni*; red circles represent *C. coli* clade 1, composed of 692 ST-828, 12 ST-1150 clonal complexes genomes and 35 unintrogressed clade 1 genomes; yellow circles represent 46 *C. coli* clade 2 genomes; green circles represent 60 *C. coli* clade 3 genomes. The scale represent the number of substitutions per site.

Assignment to the lineages was also done for the 657 *C. coli* genome sequences from Genbank/EMBL/PubMLST. We furthermore included 33 *C. coli* genomes from the IID1/IID2 study (genome sequences from project FS101072, kindly provided by Prof Craig Winstanley, University of Liverpool). This led to the following distribution over the lineages (Table 4):

Table 4. Assignment of *C. coli* genomes to agricultural and environmental lineages

		Clade 1				
	Total	**Cluster A (ST-828)**	**Cluster B (ST-1150)**	**Cluster C (unintrogr)**	**Clade 2**	**Clade 3**
Genbank/PubMLST	657	615	12	5	7	18
IID1/IID2 (FS101072)	33	33	0	0	0	0
This study	497	372	8	24	37	56

The distribution of the genomes over the lineages shows that one of the aims of the project was achieved, by significantly increasing the number of riparian (non-agricultural) genome sequences. The 30 genomes publicly available for Clade 1-Cluster C, and Clades 2 and 3 have been increased with 117 genome sequences.

3.5.5 Genomic diversity and distribution of isolation source over the *C. coli* lineages

To assess whether the genomic diversity of the 497 *C. coli* genome sequences from this study is comparable to the publicly available and IID1/IID2 genome sequences, we determined the

SNP-based phylogeny of the combined dataset (Figure 5), and annotated the tree with the isolation source:

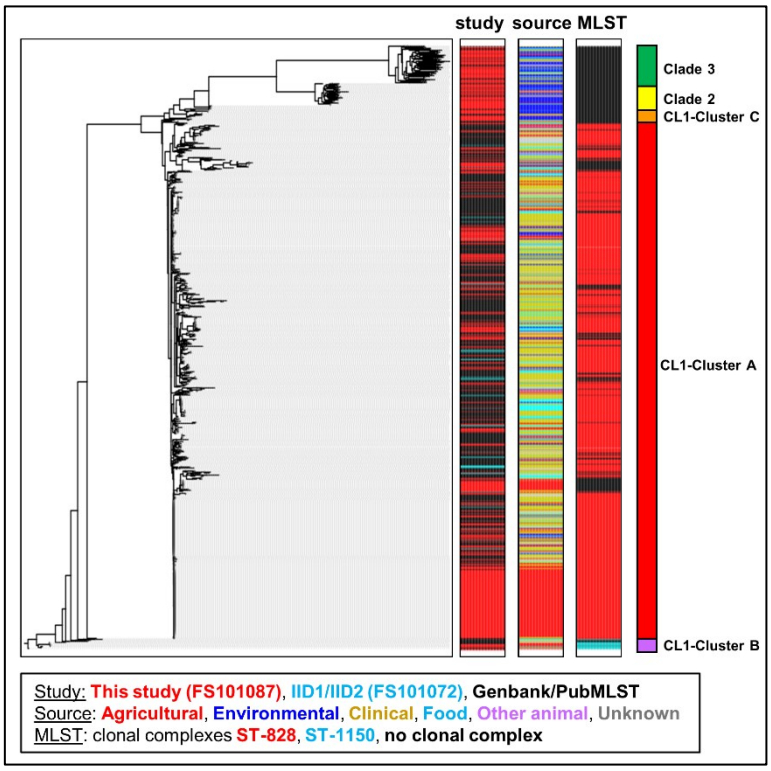


Figure 5. SNP-based phylogeny of *C. coli* isolated in this study (FS101087), the IID1/IID2 study (FS101072) and publicly available *C. coli* genome sequences (Genbank/PubMLST). The tree is shown in a rectangular format. The first column shows the origin of the sample (red, this study; light blue, IID1/IID2; black, other sources); the second column shows the isolation sources (yellow, clinical; red, agricultural; dark blue, environmental; light blue, food; purple, other animals; grey, unknown); the third column shows the MLST subdivision (red, ST-828 clonal complex; light blue, ST-1150 clonal complex; black, other MLST-types). The subdivision into clades and clusters is shown on the right hand side.

Phylogenomic diversity: The comparison of the three different groups of genome sequences (public repositories, IID1/IID2 study and this study) showed there was no clear study-specific clustering within the phylogenetic tree in the Clade 1-Cluster A (ST-828) group, suggesting that the current study FS101087 is not skewed with regard to *C. coli* genomic diversity. It also shows that this study significantly increases the number of genomes in the Clade 1-Cluster C, Clade 2 and Clade 3 lineages, in line with the aims of the project. The original observation of clearly separate lineages in the *C. coli* population holds true for this larger dataset, with no clear intermediates between the lineages. This can be clearly observed in a different visualisation of the phylogenetic tree (Figure 6).

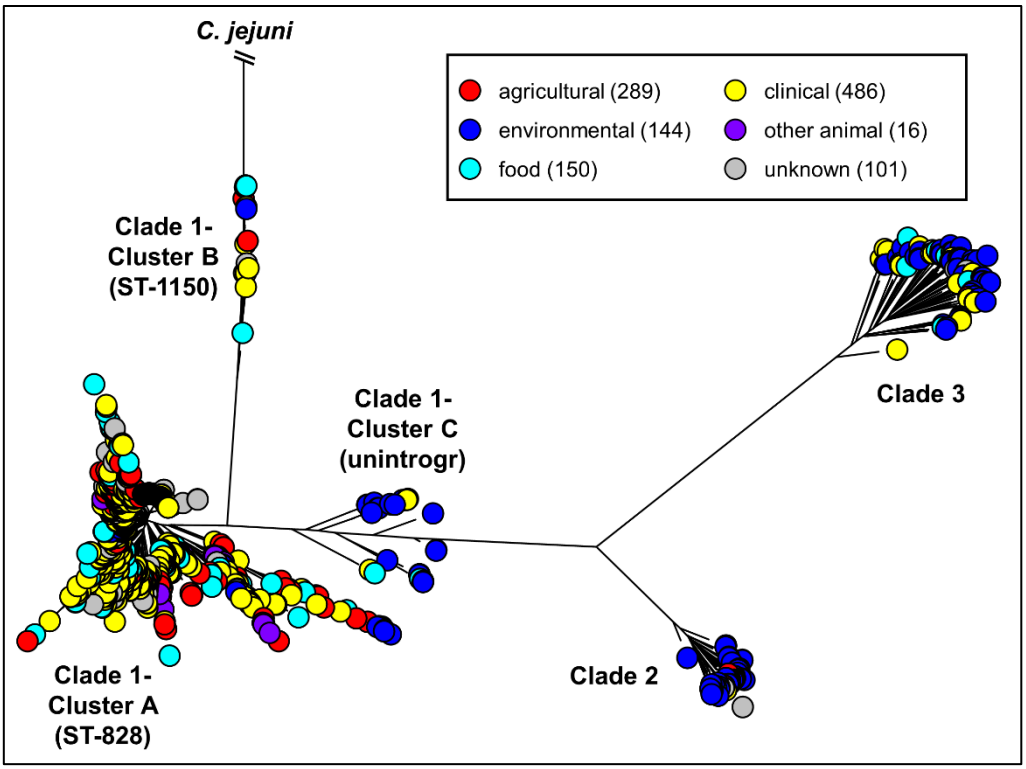
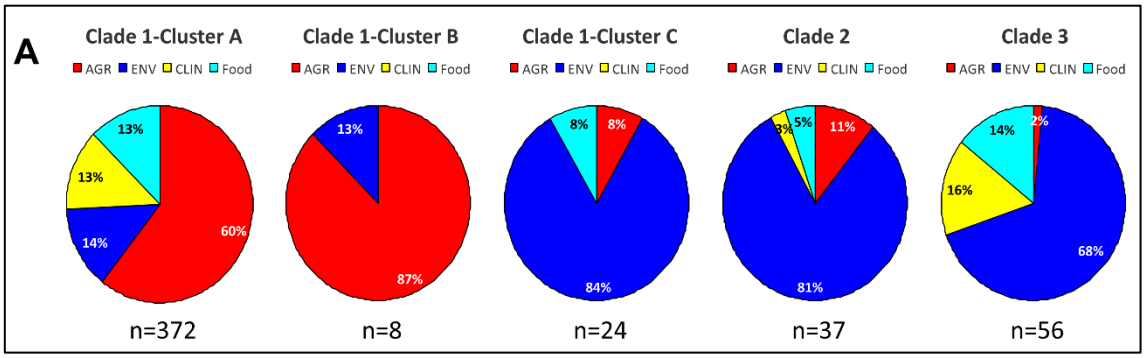


Figure 6. SNP-based phylogeny of *C. coli* isolated in this study (FS101087), the IID1/IID2 study (FS101072) and publicly available *C. coli* genome sequences (Genbank/PubMLST). The tree is shown in a radial format, with the samples colored according to isolation source.

Isolation source: The comparison of isolation sources and the five lineages (Figure 6) shows that clinical isolates and agricultural isolates are primarily found in Clade 1-Clusters A and B, whereas non-agricultural (environmental/riparian) isolates primarily cluster in Clades 1-Cluster C and Clade 2 (Figure 6). Clade 3 contains a mixture of clinical and environmental isolates. Overall this supports the suggested epidemiological link between agriculture and human disease for Clade 1-Clusters A and B isolates (16, 49, 50), whereas the mixture of environmental, food and clinical isolates in Clade 3 suggests an environmental source or non-agricultural food source for Clade 3-caused human disease. A further subdivision for the samples from this study (FS101087) is shown in Table 5 and Figure 7.



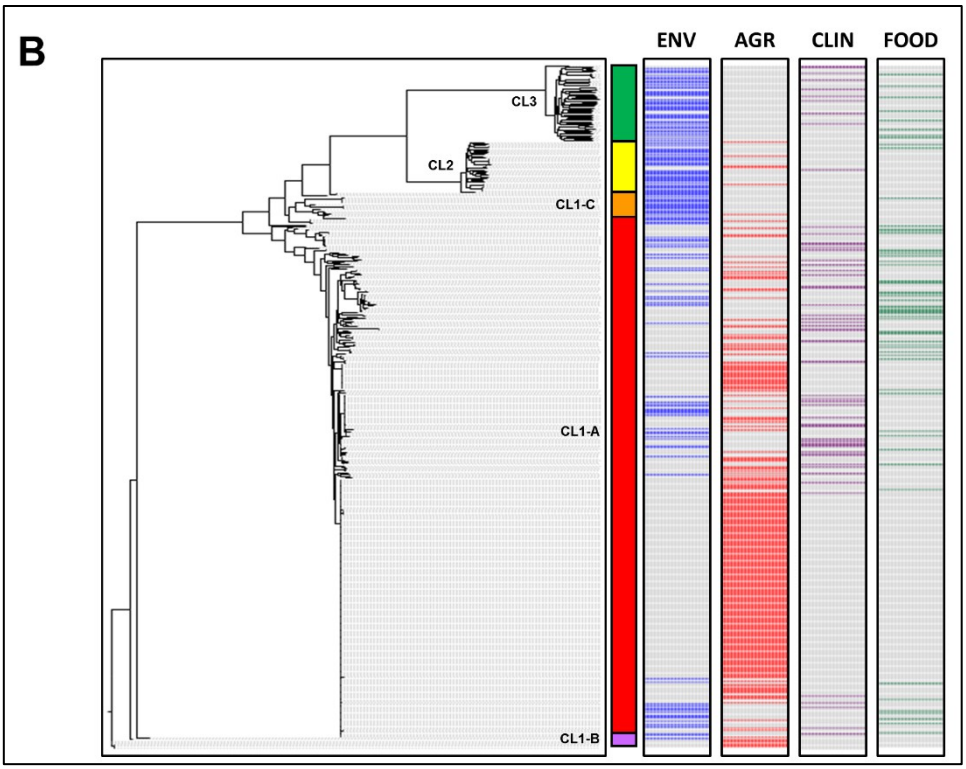


Figure 7. Distribution of isolation source in the 497 *C. coli* genomes included in this study. Panel A shows piecharts per Clade/Cluster, subdivided in agricultural sources (red), environmental sources (dark blue), clinical isolates (yellow) and food isolates (light blue). Panel B shows the distribution of these isolates in the SNP-based phylogeny. For presentational purposes, clinical isolates are shown using purple color, and food isolates using green.

Table 5. Distribution of isolates and sources over *C. coli* lineages

Lineage	Agricultural	Environmental	Clinical	Food	Total
Clade 1-Cluster A	**225**	51	49	47	**372**
Clade 1-Cluster B	**7**	1	0	0	**8**
Clade 1-Cluster C	2	**20**	0	2	**24**
Clade 2	4	**30**	1	2	**37**
Clade 3	1	**38**	9	8	**56**
Total	**239**	**140**	**59**	**59**	**497**

Yellow labelling of the cell shows the group with the highest proportion of isolates per lineage.

For the environmental samples in Clade 1-Cluster C, and Clades 2 and 3, the primary sources (based on the BIGSdb categories for source) are:

- Clade 1-Cluster C: duck (12/21), environmental waters (7/21)
- Clade 2: environmental waters (18/31), duck (6/31), soil (6/31)
- Clade 3: environmental waters (37/41), soil (3/41)

3.5.6 Comparative genomics and genome analyses

The 497 *C. coli* genome sequences have been used for an array of comparative genomics analysis techniques, with the aim to identify specific genes/functions differentiating the clades, to further epidemiological and functional analyses.

3.5.6.1 Clade 3 genomes are significantly smaller

Comparison of the genome sizes included in this study showed them to have an average size of 1,704,774 ± 78,169 bp, well within the range of available genome sequences. The smallest genome was 1,517,604 bp, the largest genome 2,015,170 bp. There was a clear distribution of genome size in each lineage (Figure 8), with Clade 3 genomes being significantly smaller than the other lineages.

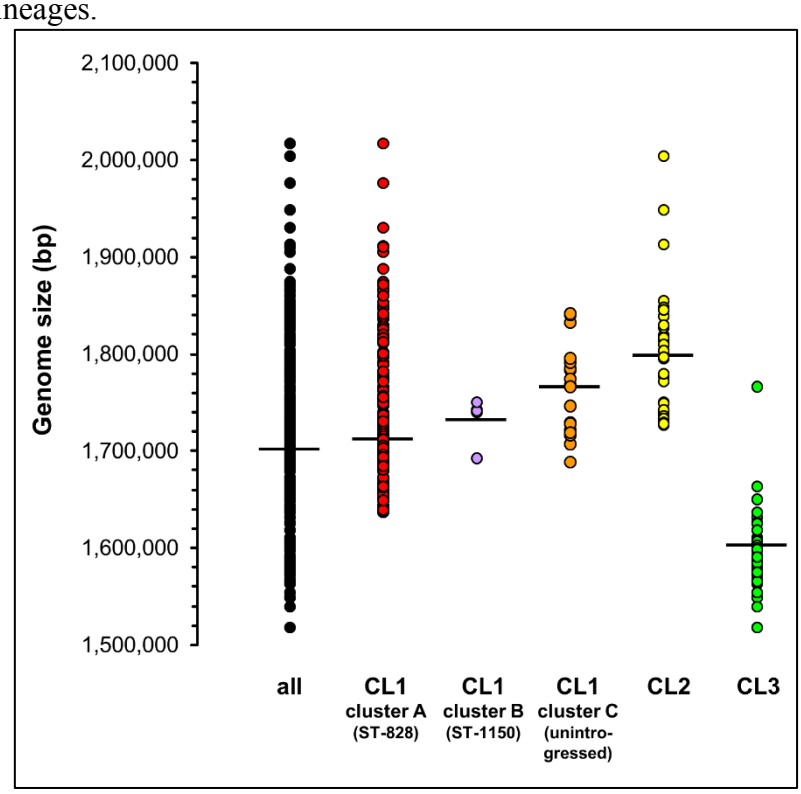


Figure 8. Distribution of genome sizes per *C. coli* lineage. Each marker represents a *C. coli* genome, the lines indicate the average genome size per group (1.70 ± 0.11 Mbp for all *C. coli*, 1.71 ± 0.06 for Clade 1-Cluster A, 1.73 ± 0.02 Mbp for Clade 1-Cluster B, 1.76 ± 0.05 Mbp for Clade 1-Cluster C, 1.79 ± 0.07 Mbp for Clade 2, and 1.60 ± 0.04 Mbp for Clade 3.

Size differences between *Campylobacter* genomes are known to be caused by the presence/absence of plasmids and by insertion elements (51, 52). Hence the genomes were searched for the presence of these plasmids and insertion elements (Table 6, Supplementary Table S5). There is a very low prevalence of mobile elements and plasmids in Clade 3 genomes, which could in part explain the smaller genome size. It could also represent adaptation to a different environmental niche.

Table 6. Distribution of mobile genetic elements and plasmids in *C. coli* lineages

Lineage	CJIE1	CJIE2	CJIE3	CJIE4	pVir	pTet	pCC31
Clade 1-Cluster A [n=372]	15.6	8.3	5.9	4.3	1.6	18.0	29.3
Clade 1-Cluster B [n=8]	75.0	100.0	none	none	none	12.5	none
Clade 1-Cluster C [n=24]	4.2	25.0	100.0	none	none	12.5	8.3
Clade 2 [n=37]	10.8	29.7	100.0	none	none	8.1	2.7
Clade 3 [n=56]	none	7.1	none	none	1.8	1.8	none

Numbers shown are the percentage of isolates positive for the mobile element/plasmid.

3.5.6.2 Distribution of variable genes and gene clusters

Previous studies on genomic variation in primarily *C. jejuni* but also *C. coli* have reported on

gene clusters showing variable distribution, and have in specific cases suggested these play a role in transmission or virulence (40, 53). We assessed the distribution of 9 of these genes/clusters representing virulence factors, metal uptake, carbon metabolism and respiration functions. Table 7 summarizes the findings, fully presented in Supplementary Table S4.

Table 7. Distribution of variable genes/gene clusters over the *C. coli* lineages

	Clade 1				
Gene / gene cluster	**Cluster A** (n=372)	**Cluster B** (n=8)	**Cluster C** (n=24)	**Clade 2** (n=37)	**Clade 3** (n=56)
DMSO reductase	none	none	none	none	100.0
lactate oxidase (*cj1585c*)	100.0	100.0	100.0	100.0	none
fucose operon	87.1	100.0	none	none	5.4
ggt carbon metabolism	none	none	none	29.7	89.3
vitamine B5 operon	100.0	100.0	100.0	100.0	100.0
Type VI secretion system	12.6	none	20.8	56.8	none
flagellar modification	87.1	87.5	70.8	62.2	42.9
serine protease (*cj1365c*)	0.8	100.0	none	none	100.0
iron uptake (*cfrA*)	100.0	100.0	100.0	100.0	53.6

Numbers shown are the percentage of isolates positive for the gene/gene cluster.

This analysis shows that there are lineage-specific genes/gene clusters. For instance, the DMSO reductase genes *dmsABCD* are restricted to Clade 3, whereas the lactate oxidase gene *cj1585c* is absent from Clade 3. Further analysis shows that these genes are located in the same region of the genome. A similar mutually exclusive distribution has been observed in *C. jejuni* (van Vliet AHM, unpublished observations). The serine protease gene *cj1365c* is absent from Clades 1-Clusters A and C, and Clade 2, but present in all Clade 1-Cluster B and Clade 3 isolates. The fucose utilisation operon is primarily found in the agricultural Clade 1-Clusters A and B, whereas the gamma-glutamyl transpeptidase GGT is restricted to Clade 2 and Clade 3 isolates. Of the virulence factors, the Type VI secretion system (54, 55) is found in over half of the Clade 2 isolates, but its prevalence is low to none in the other clades. Finally, the vitamine B5 biosynthesis operon linked with chicken colonisation in *C. jejuni* (40) is present in all *C. coli* isolates investigated.

3.5.6.3 Analysis of core and accessory genome variation in *C. coli*

To characterise core and accessory genome variation, a previously published list of all genes present in 7 reference genomes of *C. jejuni* and *C. coli* was used (10). Briefly, the list of all coding sequences (CDS) found in all 7 reference genomes was filtered using BLAST to keep only the CDS uniquely present in the dataset, removing all alleles and duplicates for each gene. The subsequent list contained 2335 genes, which are virtually contained reference sequences for all genes present in 7 reference strains, which provides a good starting point for core and accessory genome analysis, as we have published in 2014 (10). All genes from this list were used to scan 853 *C. coli* genomes included for pan-genome MLST, comprising genomes sequenced in this study as well as genomes published in other studies (10, 16).

All *C. jejuni* and *C. coli* isolates shared 622 genes from the reference pan-genome used for comparison. A total of 672 genes were shared by all *C. coli*, and more than the double (1,360 genes) were shared by all *C. jejuni*. This difference can likely be explained by the highest genetic diversity across all *C. coli* clades, compared to *C. jejuni*. A total of 59 genes were present only in *C. jejuni* and absent in *C. coli*, while 7 genes were present in *C. coli* but absent in *C. jejuni*.

3.5.6.3 Identification of regions of variation in the genome suitable for epidemiological analysis

As highlighted in a published study from a selection of these genomes, the comparison of patterns of gene presence/absence in the pan-genome genes of 192 *C. jejuni* and *C. coli* genomes was performed to identify genes that segregated by species or lineage. Segregation was either: complete, with genes present in one group and absent in the other; or frequency dependent, where genes were significantly more frequent in one lineage. Consistent with the aim of discovering informative epidemiological markers, we focused on accessory genes that were the most specific to each of the 7 examined lineages of *C. coli* and *C. jejuni*. Of the 3,933 genes of the pan-genome, there were 20 lineage-specific genes (10). Forty-eight genes were found to be differentially present in the species *C. jejuni* and *C. coli* (10). It is interesting to note that the genes present in *C. coli* (62/62, 100%) but not *C. jejuni* (0/130, 0%) are all present in the *C. coli* 76339 reference genome used to compile the pan-genome, and none were present in the *C. jejuni* reference strain NCTC 11168. This highlights the fact that if a typical single strain reference approach, based on a NCTC 11168, had been used to identify genetic markers specific to *C. coli*, all of these genes would have been missed. Twenty-seven genes were found to be present in all *C. jejuni* (130/130, 100%) and absent in all *C. coli* (0/62, 0%).

There were several genes that segregated according to 3-clade structure in *C. coli* (10, 20). One gene (*G157_03450*), annotated as encoding a reductase involved in fatty acid biosynthesis, was present in all *C. coli* clade 1 isolates (47/47) but was absent from isolates in the other *C. coli* clades (0/9) and from *C. jejuni* (0/130). Similarly, one gene (*Cc76339_10830*), encoding a hypothetical protein was present in *C. coli* clade 2 (4/4) but absent elsewhere (0/188). Three genes were present in all 5 genomes of *C. coli* clade 3 and absent in all other *C. coli* (0/57) and most *C. jejuni* (2/130) except for 2 environmental isolates. These genes putatively encoded a biotin sulfoxide reductase, a secreted serine protease and a cytochrome C-type periplasmic protein. *Campylobacter* uses short-chained fatty acids as nutrients, which are typical by-products of acetate and lactate metabolism by many gastrointestinal bacteria. One of the β-oxidation by-products of fatty acids chains is metabolized via the methylcitrate cycle. Interestingly, 4 genes specifically found in *C. coli* and not *C. jejuni* were involved in the methylcitrate cycle (10), which could highlight a preference or enhanced ability of *C. coli* strains to grow on odd-chained fatty acids compared to *C. jejuni*. With more focused development, this observation could potentially lead to the development of specific fatty-acid-rich media designed to discriminate more efficiently between *C. jejuni* and *C. coli*, or to improve isolation frequency of *C. coli* in the laboratory.

3.5.6.4 Analysis of core/accessory genome and identification of new lineage-specific genes

We also used a *de novo* screening method for identification of lineage-specific genes, based on the annotation of genome sequences using Prokka (31), and the pan/core/accessory genome analysis tools in the Roary package (37). Briefly, the annotated features were translated to amino acid sequences, grouped using CD-hit, and subsequently compared using pairwise BLAST analysis. We used relaxed criteria (80% identity at amino acid level, core if present ≥ 80% of isolates per lineage) to avoid allelic differences influencing the analyses. Analysis of a total of 4,236 genes from the 497 *C. coli* genomes in this study identified 1,369 core genes (>95% present in all 497 genomes). A total of 2,473 genes were present in <15% of genomes.

A total of 86 putative lineage-specific genes were identified from this dataset, and tested against the full 1,020 *C. coli* genome dataset consisting of genomes from this study (497), the FS101072 IID1/IID2 study (33) and the 657 Genbank/PubMLST *C. coli* genomes. A subset of 20 genes showed strong Clade/Cluster-specificity, and was further analysed agains a 4,064 *C. jejuni* genome sequence dataset as well, to test for specificity for *C. coli*. The gene presence/absence analysis for the 20 genes in the FS101087 *C. coli* genomes is presented in Supplementary Table S7, while Table 8 shows the percentage of genomes containing the 20 lineage-specific genes and their putative functions.

While many genes show lineage-specific distribution, there are very few genes which are completely limited to a specific lineage, suggesting some level of genetic transfer between lineages, possibly via natural transformation from other *C. coli* isolates or via exchange of *C. jejuni* sequences present in the same environmental niche.

3.5.6.5 Antimicrobial resistance predictions

The advent of high-throughput genomics now allows for prediction of antimicrobial resistance in isolates, including *C. coli* (11, 56-60). We have searched for the presence of 21 genes associated with resistance to aminoglycosides, β-lactams and tetracycline, and for mutations known to confer resistance to quinolones and macrolides. We did not detect any resistance to aminoglycosides in the 497 *C. coli* genomes analysed, but the other resistance markers were observed. Table 9 summarizes the abundance of these markers in the different lineages, with the detailed data provided in Supplementary Table S6.

Table 9. Distribution of putative antimicrobial resistance markers over the *C. coli* lineages

	Clade 1				
Gene (resistance to)	**Cluster A** (n=372)	**Cluster B** (n=8)	**Cluster C** (n=24)	**Clade 2** (n=37)	**Clade 3** (n=56)
aad9 (spectinomycin)	0.3	none	none	none	none
lnuC (lincomycin)	0.5	none	none	none	none
antA (streptomycin)	9.4	none	50.0	54.0	12.5
tetO (tetracycline)	19.6	25.0	4.2	none	none
bla-OXA61 (β-lactams)	78.0	100.0	4.2	13.5	30.4
23S rRNA $A_{2074}G$ (macrolides)	4.0	none	none	none	none
23S rRNA $A_{2075}G$ (macrolides)	none	none	none	none	none
gyrA $T_{86}I$ (quinolones)	12.6	none	none	none	none

Numbers shown are the percentage of isolates positive for the gene/gene cluster.

Antimicrobial resistance markers are mostly restricted to Clades 1a and 1b, with the exception of the *antA* streptomycin resistance marker, and the *bla*-OXA61 gene. However, the link between the *bla*-OXA61 gene and high-level β-lactam resistance is controversial (61). With regard to the *antA* gene, this gene was only identified recently (41) and its role has only been investigated in a small number of isolates and hence it is difficult to draw conclusions about the effect of the *antA* gene of non-agricultural clades on antimicrobial resistance. With regard to macrolide resistance, the A2075G mutation known to confer high-level macrolide resistance was absent, although the A2074G mutation was detected in a few isolates. Overall, these data suggest that antimicrobial resistance is primarily found in agricultural isolates, without a reservoir of these resistance markers in the environmental isolates.

Table 8. Distribution of lineage-specific genes in *C. coli* lineages

gene [a]	*C. coli* [b]	Clade	Annotation [c]	Clade 1 [d]			Clade 2	Clade 3	*C.jejuni*
				Clu. A	Clu. B	Clu. C			
N149_0172	15-537360	Clade 1-Cluster A	hypothetical protein	**76.5**	0.0	0.0	11.4	0.0	0.0
N149_0173	15-537360	Clade 1-Cluster A	hypothetical protein	**76.3**	0.0	0.0	2.3	0.0	1.3
N149_1346	15-537360	Clade 1-Cluster A	hypothetical protein	**99.8**	15.8	**100.0**	0.0	0.0	0.0
YSS_00925	RM4661	Clade 1-Cluster B	DUF336 domain-containing protein	**85.3**	**100.0**	3.4	0.0	0.0	0.7
YSS_01135	RM4661	Clade 1-Cluster B	phosphoesterase	**72.1**	**78.9**	0.0	13.6	0.0	0.0
WP_059007146	generic	Clade 1-Cluster C	hypothetical protein	0.0	0.0	**89.7**	0.0	0.0	0.0
WP_059007153	generic	Clade 1-Cluster C	hypothetical protein	0.0	0.0	**86.2**	2.3	0.0	0.0
WP_059007648	generic	Clade 1-Cluster C	hypothetical protein	1.8	0.0	**93.1**	0.0	0.0	0.0
WP_038848989	generic	Clade 2	hypothetical protein	0.0	0.0	3.4	**97.7**	0.0	0.0
WP_058976849	generic	Clade 2	hypothetical protein	0.0	0.0	3.4	**97.7**	0.0	0.0
WP_058977712	generic	Clade 2	hypothetical protein	0.0	0.0	0.0	**93.2**	0.0	0.0
WP_038840639	generic	Clade 2	hypothetical protein	0.0	0.0	3.4	**93.2**	0.0	0.0
BN865_04060	76339	Clade 3	biotin sulfoxide reductase	0.0	0.0	0.0	0.0	**100.0**	1.2
BN865_04070	76339	Clade 3	monoheme c-type cytochrome	0.0	0.0	0.0	0.0	**100.0**	1.2
BN865_07680c	76339	Clade 3	serine protease, peptidase S8 family	0.0	0.0	0.0	0.0	**100.0**	0.0
BN865_12110c	76339	Clade 3	hypothetical protein, autotransporter	0.0	0.0	0.0	0.0	**94.6**	0.0
BN865_12600	76339	Clade 3	hypothetical protein	0.0	0.0	0.0	0.0	**94.6**	0.0
BN865_12620	76339	Clade 3	filamentous haemagglutinin protein	0.0	0.0	0.0	0.0	**94.6**	0.0
BN865_12640	76339	Clade 3	hypothetical protein	0.0	0.0	0.0	0.0	**94.6**	0.0
BN865_01820	76339	Clade 3	bacteriohemerythrin	0.0	0.0	0.0	0.0	**98.6**	0.0

a) Genes/Genbank accession numbers for representatives for retrieval of sequences

b) *C. coli* genome for which the representative is included. Generic: no specific genome indicated in Genbank.

c) Provisional annotation

d) Numbers shown are percentages of positive genomes. Number of genomes per category: Clade 1-Cluster A: 1,020; Clade 1-Cluster B: 19; Clade 1-Cluster C: 24; Clade 2: 44; Clade 3: 74; *C. jejuni*: 4,064. Cells with yellow background highlight >70% of genomes positive for the gene.

Antimicrobial resistance markers are mostly restricted to Clades 1a and 1b, with the exception of the *antA* streptomycin resistance marker, and the *bla*-OXA61 gene. However, the link between the *bla*-OXA61 gene and high-level β-lactam resistance is controversial (61). With regard to the *antA* gene, this gene was only identified recently (41) and its role has only been investigated in a small number of isolates and hence it is difficult to draw conclusions about the effect of the *antA* gene of non-agricultural clades on antimicrobial resistance. With regard to macrolide resistance, the A2075G mutation known to confer high-level macrolide resistance was absent, although the A2074G mutation was detected in a few isolates. Overall, these data suggest that antimicrobial resistance is primarily found in agricultural isolates, without a reservoir of these resistance markers in the environmental isolates.

3.5.7 Database submissions

The Illumina sequencing reads, genome assemblies, strain codes and metadata have been uploaded to the European Nucleotide Archive (ENA) at the European Bioinformatics Institute (EBI) as project PRJEB11972, with accession number FAXG01-FAXH01, FAXQ01-FBQZ01 and CP006702-CP006703 for the assembled genomes, and to the Campylobacter PubMLST database (IDs 32298 and 43935-44440). A full list of the specific EBI accession numbers and PubMLST IDs is given in Supplementary Table S2. A provisional annotation of the genome sequences is available via http://www.ifr.ac.uk/campylobacter/campylobactercoligenomics.html

3.6 OVERVIEW OF PROJECT TASKS

Tasks 01/01, 01/03 Isolation of genomic DNA of *C. coli* isolates
Tasks 01/03, 01/05: Sequencing and genome assembly of *C. coli* isolates
These tasks have been successfully completed (see sections 3.5.1 and 3.5.2).

Task 02/01: Draft annotation of *C. coli* genome sequences
Task 02/02: Identification of core and accessory genomes of *C. coli* and comparative genomics
Task 02/03: identification of regions of variation in the genome suitable for epidemiological analysis
These tasks were successfully completed (see sections 3.5.4, 3.5.5 and 3.5.6.1 to 3.5.6.4).

Task 03/01: Linking genome sequences with MLST data and other metadata
Task 03/02: linking of genome sequences with phenotypic typing methods
These tasks were successfully completed (see section 3.5.3 and 3.5.6.5).

Task 04/01: Release of curated database of *Campylobacter coli* genome sequences
This task was successfully completed (see section 3.5.7).

4. CONCLUSIONS

As with many large genome sequencing projects, this project has generated a substantial amount of data, which are actively being analysed. This report contains the first batch of analyses, led by the tasks described in the original proposal to FSA. Further in-depth analyses will follow and will be reported in peer-reviewed publications, which will acknowledge the funding from FSA.

Specific conclusions from this project are:

1. We have determined and annotated the draft genome sequence of 497 *C. coli* isolates and 10 *C. jejuni* isolates. The *C. jejuni* isolates were previously misclassified as *C. coli*. Of these 497 *C. coli* genomes, five were completed using PacBio sequencing for use as reference sequence.
2. *C. jejuni* and *C. coli* are clearly distinct at the level of the genome sequence, consistent with earlier reports.
3. The genome sequence-based population structure of *C. coli* is highly structured with five distinct lineages identified, consistent with earlier reports.
4. Agricultural and clinical isolates of *C. coli* are found primarily in Clade 1-Cluster A (clonal complex ST-828) and Cluster B (clonal complex ST-1150), whereas environmental isolates are primarily found in Clade 1-Cluster C (unintrogressed), Clade 2 and Clade 3.
5. The genome sequences were successfully used to predict antimicrobial susceptibility profiles, showing that antimicrobial resistance was mostly restricted to agricultural isolates, and that genes/mutations conferring resistance to aminoglycosides, lincosamides and macrolides were rare to absent. The environmental Clade 1-Cluster C, and Clades 2 and 3 contained very few resistance markers, suggesting that these are probably not an important reservoir for antimicrobial resistance.
6. The clear separation of Clade 1-Cluster C, Clade 2 and Clade 3 from Clade 1-Clusters A and B, and from each other suggests that these clades may have different environmental niches, which is in part supported by the source of isolation reported for these lineages.
7. Candidate genes and sequences specific for one or more *C. coli* lineages have been identified that may be used for epidemiological purposes.

5. RECOMMENDATIONS

1. This project highlights the importance of open-ended investigations with regard to genome sequencing. While monitoring of clinical isolates, such as the excellent FSA-supported Oxfordshire Campylobacter survey, gives a richess of information on the gene pool and population structure of clinical isolates, it cannot be properly appreciated without the comparator of environmental and agricultural isolates which represent the other facets of the pathogens' lifestyle. Hence it is strongly recommended that future large sequencing projects include a rationale for isolate selection and how comparator material will be obtained, and develop a clear template for minimally required epidemiological information.
2. The genome sequences determined in this project provide a baseline for assessing the role of agricultural and non-agricultural (riparian/environmental) sources in causing foodborne illness in the UK. The lack of agricultural isolates in Clade 1-Cluster C, and Clades 2 and 3, versus the low prevalence of environmental isolates in Clade 1-Clusters A and B, suggests that there is a physical or other barrier preventing mixing of these populations.
3. It is important to (financially) support a mechanism where *Campylobacter* genome sequences and annotations can be accessed and curated. Currently this is scattered over initiatives such as the Campylobacter PubMLST database, the public Genbank and EMBL databases, and other initiatives. The Campylobacter PubMLST database and resources such as EnteroBase (https://enterobase.warwick.ac.uk) and MRC Climb (http://www.climb.ac.uk) can assist the scientific community and regulators, and hopefully prevent fragmentation of information and datasets.
4. Future studies with the datasets will include genome-wide association studies (GWAS) such as shown in section 3.5.6.4, and utilise genome-based MLST schemes and other phylogenetic clustering methods to identify markers that may explain the genetic differences between the lineages and may provide a biological explanation for the different environmental niches occupied. Any publications including the genomic data from this study will specifically acknowledge the funder FSA and project number FS101087.
5. The four completed genomes for Clades 1-Cluster C, Clades 2 and 3 generated for this study are to date the only complete genomes available for these clades, and have been provided to the scientific community for use in improved mapping and assembly of short-read WGS data for epidemiological purposes, and for future studies on the pathogenicity and lifestyle of environmental *C. coli* isolates.

6. ACKNOWLEDGEMENTS

We would like to thank the Food Standards Agency for their funding of this project (FSA Project FS101087), BBSRC for IFRs core funding, MRC for funding of CLIMB at the University of Swansea, and the Environmental & Social Ecology of Human Infectious Diseases Initiative (ESEI)-funded ENIGMA study from which isolates were obtained. We thank the sequencing staff at The Genome Analysis Centre for Illumina and PacBio sequencing, staff at Public Health England and University of Liverpool, Clare Barker (PHE/University of Oxford) for uploading genome sequences in PubMLST, Pippa Connerton (University of Nottingham) and Lisa Williams (University of Bristol) for sending *C. coli* isolates, and Prof Craig Winstanley and Dr Sam Haldenby (University of Liverpool) for sending *C. coli* genome sequences from the IID1/IID2 studies. This report made use of the PubMLST website (http://pubmlst.org/) developed by Keith Jolley and cited at the University of Oxford. The development of that website was funded by the Wellcome Trust. We also thank Martin Maiden, Keith Jolley and their staff, and the contributors of genome sequences available from the Campylobacter pubMLST website.

7. REFERENCES

1. Nichols GL, Richardson JF, Sheppard SK, 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.
2. Tam CC, O'Brien SJ, Tompkins DS, Bolton FJ, Berry L, Dodds J, Choudhury D, Halstead F, Iturriza-Gomara M, Mather K, Rait G, Ridge A, Rodrigues LC, Wain J, Wood B, Gray JJ (2012) Changes in causes of acute gastroenteritis in the United Kingdom over 15 years: microbiologic findings from 2 prospective, population-based studies of infectious intestinal disease. Clin Infect Dis **54**, 1275-1286.
3. Tam CC, Rodrigues LC, Viviani L, Dodds JP, Evans MR, Hunter PR, Gray JJ, Letley LH, Rait G, Tompkins DS, O'Brien SJ (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.
4. Doorduyn Y, Van Den Brandhof WE, Van Duynhoven YT, Breukink BJ, Wagenaar JA, Van Pelt W (2010) Risk factors for indigenous Campylobacter jejuni and Campylobacter coli infections in The Netherlands: a case-control study. Epidemiol Infect **138**, 1391-1404.
5. Sheppard SK, Colles FM, McCarthy ND, Strachan NJ, Ogden ID, Forbes KJ, Dallas JF, Maiden MC (2011) Niche segregation and genetic structure of Campylobacter jejuni populations from wild and agricultural host species. Mol Ecol **20**, 3484-3490.
6. Sheppard SK, McCarthy ND, Falush D, Maiden MC (2008) Convergence of Campylobacter species: implications for bacterial evolution. Science **320**, 237-239.
7. Cody AJ, McCarthy ND, Jansen van Rensburg M, Isinkaye T, Bentley SD, Parkhill J, Dingle KE, Bowler IC, Jolley KA, Maiden MC (2013) Real-time genomic epidemiological evaluation of human Campylobacter isolates by use of whole-genome multilocus sequence typing. J Clin Microbiol **51**, 2526-2534.
8. Cody AJ, McCarthy NM, Wimalarathna HL, Colles FM, Clark L, Bowler IC, Maiden MC, Dingle KE (2012) A longitudinal 6-year study of the molecular epidemiology of clinical campylobacter isolates in Oxfordshire, United kingdom. J Clin Microbiol **50**, 3193-3201.
9. Sheppard SK, Jolley KA, Maiden MC (2012) A Gene-By-Gene Approach to Bacterial Population Genomics: Whole Genome MLST of Campylobacter. Genes **3**, 261-277.
10. Meric G, Yahara K, Mageiros L, Pascoe B, Maiden MC, Jolley KA, Sheppard SK (2014) A reference pan-genome approach to comparative bacterial genomics: identification of novel epidemiological markers in pathogenic Campylobacter. PloS one **9**, e92798.
11. Zhao S, Tyson GH, Chen Y, Li C, Mukherjee S, Young S, Lam C, Folster JP, Whichard JM, McDermott PF (2015) Whole-Genome Sequencing Analysis Accurately Predicts Antimicrobial Resistance Phenotypes in Campylobacter spp. Appl Environ Microbiol **82**, 459-466.
12. Sheppard SK, Dallas JF, Strachan NJ, MacRae M, McCarthy ND, Wilson DJ, Gormley FJ, Falush D, Ogden ID, Maiden MC, Forbes KJ (2009) Campylobacter genotyping to determine the source of human infection. Clin Infect Dis **48**, 1072-1078.
13. Sheppard SK, Dallas JF, MacRae M, McCarthy ND, Sproston EL, Gormley FJ, Strachan NJ, Ogden ID, Maiden MC, Forbes KJ (2009) Campylobacter genotypes from food animals, environmental sources and clinical disease in Scotland 2005/6. Int J Food Microbiol **134**, 96-103.
14. Lang P, Lefebure T, Wang W, Pavinski Bitar P, Meinersmann RJ, Kaya K, Stanhope MJ (2010) Expanded multilocus sequence typing and comparative genomic hybridization of Campylobacter coli isolates from multiple hosts. Appl Environ Microbiol **76**, 1913-1925.
15. Lefebure T, Bitar PD, Suzuki H, Stanhope MJ (2010) Evolutionary dynamics of complete Campylobacter pan-genomes and the bacterial species concept. Genome Biol Evol **2**, 646-655.

16. Sheppard SK, Didelot X, Jolley KA, Darling AE, Pascoe B, Meric G, Kelly DJ, Cody A, Colles FM, Strachan NJ, Ogden ID, Forbes K, French NP, Carter P, Miller WG, McCarthy ND, Owen R, Litrup E, Egholm M, Affourtit JP, Bentley SD, Parkhill J, Maiden MC, Falush D (2013) Progressive genome-wide introgression in agricultural Campylobacter coli. Mol Ecol **22**, 1051-1064.
17. Pearson BM, Rokney A, Crossman LC, Miller WG, Wain J, van Vliet AH (2013) Complete Genome Sequence of the Campylobacter coli Clinical Isolate 15-537360. Genome Announc **1**, e01056-01013.
18. Gurtler M, Alter T, Kasimir S, Fehlhaber K (2005) The importance of Campylobacter coli in human campylobacteriosis: prevalence and genetic characterization. Epidemiol Infect **133**, 1081-1087.
19. Rivoal K, Ragimbeau C, Salvat G, Colin P, Ermel G (2005) Genomic diversity of Campylobacter coli and Campylobacter jejuni isolates recovered from free-range broiler farms and comparison with isolates of various origins. Appl Environ Microbiol **71**, 6216-6227.
20. Sheppard SK, Dallas JF, Wilson DJ, Strachan NJ, McCarthy ND, Jolley KA, Colles FM, Rotariu O, Ogden ID, Forbes KJ, Maiden MC (2010) Evolution of an agriculture-associated disease causing Campylobacter coli clade: evidence from national surveillance data in Scotland. PloS one **5**, e15708.
21. Snipen L, Wassenaar T, Altermann E, Olson J, Kathariou S, Lagesen K, Takamiya M, Knochel S, Ussery DW, Meinersmann RJ (2012) Analysis of evolutionary patterns of genes in Campylobacter jejuni and C. coli. Microb Inform Exp **2**, 8.
22. Skarp-de Haan CP, Culebro A, Schott T, Revez J, Schweda EK, Hanninen ML, Rossi M (2014) Comparative genomics of unintrogressed Campylobacter coli clades 2 and 3. BMC genomics **15**, 129.
23. Linton D, Lawson AJ, Owen RJ, Stanley J (1997) PCR detection, identification to species level, and fingerprinting of Campylobacter jejuni and Campylobacter coli direct from diarrheic samples. J Clin Microbiol **35**, 2568-2572.
24. Wilson K (2001) Preparation of genomic DNA from bacteria. Curr Protoc Mol Biol **Chapter 2**, Unit 2 4.
25. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol **19**, 455-477.
26. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res **18**, 821-829.
27. Magoc T, Pabinger S, Canzar S, Liu X, Su Q, Puiu D, Tallon LJ, Salzberg SL (2013) GAGE-B: an evaluation of genome assemblers for bacterial organisms. Bioinformatics **29**, 1718-1725.
28. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics **29**, 1072-1075.
29. Darling AC, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res **14**, 1394-1403.
30. Rissman AI, Mau B, Biehl BS, Darling AE, Glasner JD, Perna NT (2009) Reordering contigs of draft genomes using the Mauve aligner. Bioinformatics **25**, 2071-2073.
31. Seemann T (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics **30**, 2068-2069.
32. Treangen TJ, Ondov BD, Koren S, Phillippy AM (2014) The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol **15**, 524.

33. van Vliet AHM, Kusters JG (2015) Use of alignment-free phylogenetics for rapid genome sequence-based typing of Helicobacter pylori virulence markers and antibiotic susceptibility. J Clin Microbiol **53**, 2877-2888.
34. Sims GE, Jun SR, Wu GA, Kim SH (2009) Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc Natl Acad Sci U S A **106**, 2677-2682.
35. Dingle KE, Colles FM, Falush D, Maiden MC (2005) Sequence typing and comparison of population biology of Campylobacter coli and Campylobacter jejuni. J Clin Microbiol **43**, 340-347.
36. Kruczkiewicz P, Mutschall S, Barker D, Thomas J, Van Domselaar G, Gannon VPJ, Carrillo CD, Taboada EN (2013) MIST: a tool for rapid in silico generation of molecular data from bacterial genome sequences. Proc Bioinform 2013, 316–323.
37. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J (2015) Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics **31**, 3691-3693.
38. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22**, 1658-1659.
39. Sheppard SK, Cheng L, Meric G, de Haan CP, Llarena AK, Marttinen P, Vidal A, Ridley A, Clifton-Hadley F, Connor TR, Strachan NJ, Forbes K, Colles FM, Jolley KA, Bentley SD, Maiden MC, Hanninen ML, Parkhill J, Hanage WP, Corander J (2014) Cryptic ecology among host generalist Campylobacter jejuni in domestic animals. Mol Ecol **23**, 2442-2451.
40. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, Bentley SD, Maiden MC, Parkhill J, Falush D (2013) Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. Proc Natl Acad Sci U S A **110**, 11923-11927.
41. Olkkola S, Culebro A, Juntunen P, Hanninen ML, Rossi M (2016) Functional genomics in Campylobacter coli identified a novel streptomycin resistance gene located in a hypervariable genomic region. Microbiology, Epub ahead of print, 5 May 2016. doi: 2010.1099/mic.2010.000304.
42. Inouye M, Dashnow H, Raven LA, Schultz MB, Pope BJ, Tomita T, Zobel J, Holt KE (2014) SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Med **6**, 90.
43. Jolley KA, Maiden MC (2010) BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics **11**, 595.
44. Sheppard SK, Jolley KA, Maiden MCJ (2012) A Gene-By-Gene Approach to Bacterial Population Genomics: Whole Genome MLST of Campylobacter. Genes **3**, 261-277.
45. Cody AJ, Maiden MJ, Dingle KE (2009) Genetic diversity and stability of the porA allele as a genetic marker in human Campylobacter infection. Microbiology **155**, 4145-4154.
46. Weerakoon DR, Borden NJ, Goodson CM, Grimes J, Olson JW (2009) The role of respiratory donor enzymes in Campylobacter jejuni host colonization and physiology. Microbial pathogenesis **47**, 8-15.
47. Abuoun M, Manning G, Cawthraw SA, Ridley A, Ahmed IH, Wassenaar TM, Newell DG (2005) Cytolethal distending toxin (CDT)-negative Campylobacter jejuni strains and anti-CDT neutralizing antibodies are induced during human infection but not during colonization in chickens. Infect Immun **73**, 3053-3062.
48. Guerry P, Poly F, Riddle M, Maue AC, Chen YH, Monteiro MA (2012) Campylobacter polysaccharide capsules: virulence and vaccines. Front Cell Infect Microbiol **2**, 7.
49. Dearlove BL, Cody AJ, Pascoe B, Meric G, Wilson DJ, Sheppard SK (2016) Rapid host switching in generalist Campylobacter strains erodes the signal for tracing human infections. ISME J **10**, 721-729.

50. Nohra A, Grinberg A, Midwinter AC, Marshall JC, Collins-Emerson JM, French NP (2016) Molecular Epidemiology of Campylobacter coli Strains Isolated from Different Sources in New Zealand between 2005 and 2014. Appl Environ Microbiol **82**, 4363-4370.
51. Fouts DE, Mongodin EF, Mandrell RE, Miller WG, Rasko DA, Ravel J, Brinkac LM, DeBoy RT, Parker CT, Daugherty SC, Dodson RJ, Durkin AS, Madupu R, Sullivan SA, Shetty JU, Ayodeji MA, Shvartsbeyn A, Schatz MC, Badger JH, Fraser CM, Nelson KE (2005) Major structural differences and novel potential virulence mechanisms from the genomes of multiple *Campylobacter* species. PLoS Biol **3**, e15.
52. Parker CT, Quinones B, Miller WG, Horn ST, Mandrell RE (2006) 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 **44**, 4125-4135.
53. Vorwerk H, Huber C, Mohr J, Bunk B, Bhuju S, Wensel O, Sproer C, Fruth A, Flieger A, Schmidt-Hohagen K, Schomburg D, Eisenreich W, Hofreuter D (2015) A transferable plasticity region in Campylobacter coli allows isolates of an otherwise non-glycolytic food-borne pathogen to catabolize glucose. Mol Microbiol **98**, 809-830.
54. Corcionivoschi N, Gundogdu O, Moran L, Kelly C, Scates P, Stef L, Cean A, Wren B, Dorrell N, Madden RH (2015) Virulence characteristics of hcp (+) Campylobacter jejuni and Campylobacter coli isolates from retail chicken. Gut Pathog **7**, 20.
55. Bleumink-Pluym NM, van Alphen LB, Bouwman LI, Wosten MM, van Putten JP (2013) Identification of a functional type VI secretion system in Campylobacter jejuni conferring capsule polysaccharide sensitive cytotoxicity. PLoS Pathog **9**, e1003393.
56. Zhao S, Mukherjee S, Chen Y, Li C, Young S, Warren M, Abbott J, Friedman S, Kabera C, Karlsson M, McDermott PF (2015) Novel gentamicin resistance genes in Campylobacter isolated from humans and retail meats in the USA. J Antimicrob Chemother **70**, 1314-1321.
57. Qin S, Wang Y, Zhang Q, Zhang M, Deng F, Shen Z, Wu C, Wang S, Zhang J, Shen J (2014) Report of ribosomal RNA methylase gene erm(B) in multidrug-resistant Campylobacter coli. J Antimicrob Chemother **69**, 964-968.
58. Wang Y, Zhang M, Deng F, Shen Z, Wu C, Zhang J, Zhang Q, Shen J (2014) Emergence of multidrug-resistant Campylobacter species isolates with a horizontally acquired rRNA methylase. Antimicrob Agents Chemother **58**, 5405-5412.
59. Delsol AA, Sunderland J, Woodward MJ, Pumbwe L, Piddock LJ, Roe JM (2004) Emergence of fluoroquinolone resistance in the native Campylobacter coli population of pigs exposed to enrofloxacin. J Antimicrob Chemother **53**, 872-874.
60. Griggs DJ, Peake L, Johnson MM, Ghori S, Mott A, Piddock LJ (2009) Beta-lactamase-mediated beta-lactam resistance in Campylobacter species: prevalence of Cj0299 (bla OXA-61) and evidence for a novel beta-Lactamase in C. jejuni. Antimicrob Agents Chemother **53**, 3357-3364.
61. Juntunen P, Heiska H, Hanninen ML (2012) Campylobacter coli isolates from Finnish farrowing farms using aminopenicillins: high prevalence of bla(OXA-61) and beta-lactamase production, but low MIC values. Foodborne Pathog Dis **9**, 902-906.

8. LIST OF APPENDICES

Supplementary Tables:

- Supplementary Table S1. Information on *C. coli* isolates included in this study (Excel format)
- Supplementary Table S2. Multi Locus Sequence Typing data for the *C. coli* isolates included in this study (Excel format)
- Supplementary Table S3: Sequencing and genome assembly statistics, and accession numbers of EBI/ENA and Campylobacter PubMLST databases (Excel format)
- Supplementary Table S4. Presence/absence analysis of genes/gene clusters known to be differentially distributed in *C. jejuni* and *C. coli* (Excel format)
- Supplementary Table S5. Presence/absence analysis of mobile genetic elements and plasmids of *C. jejuni* and *C. coli* (Excel format)
- Supplementary Table S6. Presence/absence analysis of Clade/Cluster-specific genes identified in this study (Excel format)
- Supplementary Table S7. Presence/absence analysis of antimicrobial resistance markers in *C. coli* isolates included in this study (Excel format)