INLAND NORWAY UNIVERSITY OF APPLIED SCIENCES

Faculty of Applied Ecology, Agricultural Sciences and Biotechnology

Ewelina Katarzyna Rojek

Master's thesis

Molecular typing, antimicrobial resistance profiling and phylogeny of *Campylobacter* based on whole genome sequencing

Masters in Applied and Commercial Biotechnology

2018

Consent to lending by University College Library YES ☒ NO ☐

Consent to accessibility in digital archive Brage YES ☒ NO ☐

Acknowledgement

I would like to express my special appreciation and thanks to associate professor Rafi Ahmad for giving me the wonderful opportunity to perform this project under his supervision and guidance. I would like to thank you for your time and advice provided me throughout this thesis process as well as for challenging me in bioinformatics about which I knew so little before.

I would like to acknowledge my sincere gratitude to dr. Mohammed Umaer Naseer for giving me generous guidance in microbiology, for all valuable suggestions and encouragement for the completion of this work. Thank you for critical reading, technical support and most of all for opening the doors of laboratories at FHI.

Thanks to all the professors and staff of Inland Norway University of Applied Sciences for widening my knowledge about biotechnology but also for your kindness and patience. Special thanks to Anne Bergljot Falck-Ytter and Svein Birger Wærvågen who made my studies here possible.

I also thank the staff of the National Reference Laboratory for Enteropathogenic Bacteria at FHI for help during the experimental part of this project, and for a fantastic atmosphere in the laboratory.

Lastly, I would like to thank my parents for their love and constant moral support. Thanks also to Mats for heartening, motivating and supporting me in everyday struggles.

Hamar, 17th September 2018

Ewelina K. Rojek

Abbreviations

AMR – Antimicrobial Resistance

BAM – Binary Alignment Map

BAP – Bacterial Analysis Pipeline

BLAST – Basic Local Alignment Search Tool

BWA – Burrows-Wheeler Aligner

cgMLST – core genome Multilocus Sequence Typing

CDT – Cytolethal Distending Toxin

CIP – Ciprofloxacin

DBG – *De Bruijn* graph

ERM – Erythromycin

EU – European Union

EUCAST – European Society of Clinical Microbiology and Infectious Diseases

GBS – Guillain-Barré Syndrome

GEN – Gentamycin

MFS – Miller Fisher Syndrome

MIC – Minimum Inhibitory Concentration

ML – Maximum Likelihood

MLST – Multilocus Sequence Typing

MR – Multi Resistance

MSIS – Norwegian Surveillance System for Communicable Diseases

NGS – Next-Generation Sequencing

NIPH – Norwegian Institute of Public Health

NJ – Neighbour-Joining

NX – Nalidixic Acid

OLC – Overlap-Layout-Consensus assembly paradigm

RGI – Resistance Gene Identifier

SAM – Sequence Alignment Map

SD – Standard Deviation

SNP – Single Nucleotide Polymorphisms

TET – Tetracycline

UPGMA – Unweighted Pair Group Method with Arithmetic Mean

VFDB – Virulence Factors Database

WGS – Whole Genome Sequencing

WHO – World Health Organization

Table of Contents

ABSTRACT **9**

1. INTRODUCTION **10**

1.1 BACKGROUND OF RESEARCH 10

1.2 CAMPYLOBACTER 11

1.2.1 Virulence and survival factors of Campylobacter *11*

1.2.2 Genomic diversity of Campylobacter *13*

1.2.3 Campylobacter jejuni *13*

1.2.4 Campylobacter coli *13*

1.3 EPIDEMIOLOGY OF *CAMPYLOBACTER* SPECIES 14

1.3.1 Incidence and clinical features *14*

1.3.2 Reservoirs *15*

1.3.3 Epidemiology in Norway *16*

1.3.4 Antimicrobial resistance in human Campylobacter isolates *17*

1.4 NEXT-GENERATION SEQUENCING AND GENOME ASSEMBLY 18

1.4.1 Whole genome sequencing *20*

1.4.2 Illumina Sequencing *20*

1.4.3 Genome assembly *21*

1.4.3.1 *De novo* assembly 23

1.4.3.2 Reference-based mapping 24

1.5 MOLECULAR TYPING 25

1.5.1 Core genome multilocus sequence typing *25*

1.6 PHYLOGENY 26

1.6.1 Construction of phylogenetic trees *27*

1.7 THE AIM OF THE STUDY 28

2. MATERIAL AND METHODS 29

2.1 BACTERIAL ISOLATES 29

2.2 ISOLATION OF DNA 30

2.3 ANTIMICROBIAL SUSCEPTIBILITY TESTING 31

2.4 WHOLE GENOME SEQUENCING 31

2.5 ASSEMBLY OF *CAMPYLOBACTER* GENOME 32

2.5.1 Quality control and improvement 32

2.5.2 De novo assembly 32

2.5.3 Reference-based mapping 33

2.6 PHYLOGENETIC APPROACHES OF SEQUENCED DATA 34

2.6.1 Core genome MLST (cgMLST) 34

2.6.2 Core genome tree 34

2.6.3 SNP tree 35

2.7 DETECTION OF VIRULENCE AND ANTIMICROBIAL RESISTANCE GENES BASED ON WGS DATA 35

3. RESULTS 37

3.1 ILLUMINA SEQUENCING DATA SETS 37

3.2 ASSESSMENT OF WHOLE GENOME ASSEMBLIES 38

3.2.1 De novo assembly 38

3.2.2 Reference-based mapping 41

3.3 PHYLOGENETIC ANALYSIS 43

3.3.1 cgMLST 43

3.3.2 Phylogeny based on core genome 47

3.3.3 Phylogeny based on SNP 49

3.4 WGS BASED DETECTION OF VIRULENCE AND ANTIMICROBIAL RESISTANCE GENES 53

3.4.1 Virulence genes detection 53

3.4.2 Antimicrobial resistance genes detection.....53

3.5 ANTIMICROBIAL SUSCEPTIBILITY TESTING.....55

4. DISCUSSION.....57

4.1 DOWNSTREAM ANALYSIS OF SEQUENCE DATA 57

4.2 CGMLST AND PHYLOGENY 58

4.3 ANTIMICROBIAL RESISTANCE AND VIRULENCE 59

5. CONCLUSION.....62

6. FUTURE PERSPECTIVES.....63

7. REFERENCES 64

SUPPLEMENTARY FILES 71

Abstract

Campylobacter jejuni and *Campylobacter coli* are zoonotic pathogens causing diarrhoeal illnesses. Between 2009 and 2017 campylobacteriosis affected a total of 26,676 people in Norway, mostly involving *C. jejuni*. Increased number of *Campylobacter* resistant to several antimicrobial drugs has become a public health concern, and further research into molecular mechanisms conferring resistance is needed.

In this study, whole genome sequence (WGS) data from 35 *C. jejuni* and five *C. coli* isolates were used for: i) molecular typing using the core genome multilocus sequence typing (cgMLST) methodology; ii) conduct evolutionary relationships analysis between isolates based on core genome and single nucleotide polymorphisms; iii) identification of antimicrobial resistance genotypes, to correlate to resistance phenotypes; iv) identification of cluster thresholds for outbreak detection. The higher discriminatory power of WGS revealed that two isolates which had been phenotypically classified as *C. jejuni* were, in fact, *C. coli.*

Moreover, phylogenetic analysis enabled to separate the epidemiologically related strains from those which were non-related, resulting in the identification of three clusters. Reference-based cgMLST approach with 637 shared loci established that epidemiologically linked isolates within an outbreak differed by ≤ 2 allelic variations. A similar number of genomic differences (≤ 2 SNPs) was observed by SNP-based phylogeny. Antimicrobial susceptibility testing followed by the detection of resistance genotypes revealed that 26/40 isolates were resistant to at least one antimicrobial drug and a total of 9 isolates were multidrug resistant. Mostly, antimicrobial resistance was linked to resistance genes such as *oxa-61*, homologs of *tet* and *aph*, and genes encoding cmeABC efflux pump system; however phenotypic resistance towards erythromycin, tetracycline, and gentamycin was detected within isolates that did not carry *cme*, *tet*, or *aph* genes homologs.

This study shows that WGS, could be used for phylogenetic inference using tools such as multilocus sequence typing; WGS is also useful for investigation of bacterial outbreaks and provides high resolution for purposes of surveillance, prevention, and control of bacterial illnesses.

1. Introduction

1.1 Background of research

Infections with *Campylobacter* is one of the leading bacterial causes of gastroenteritis in Norway, and worldwide. Most of the human campylobacteriosis cases are caused by *Campylobacter jejuni* (90%) and to a lesser extent *Campylobacter coli* (Cody et al., 2012). The infection manifests as inflammation, abdominal pain, fever and diarrhoea (Black, Levine, Clements, Hughes, & Blaser, 1988). In some instances, infections with *C. jejuni* result in neuropathological diseases such as Guillain-Barré syndrome (GBS), Miller Fisher syndrome (MFS) or reactive arthritis (Dingle et al., 2001). Although most of the infections are mild and self-limiting, some of them might become serious (e.g., severe/prolonged gastroenteritis, suspected septicaemia) and require antibiotic treatment. Currently, the macrolide group of antibiotics, which includes erythromycin, is the recommended treatment of campylobacteriosis (Engberg, Aarestrup, Taylor, Gerner-Smidt, & Nachamkin, 2001). However, erythromycin-resistant *Campylobacter* have been detected, and it is speculated that macrolide resistance within these bacteria might be related to unknown virulence markers (Helms, Simonsen, Olsen, & Mølbak, 2005). Antimicrobial resistance in *Campylobacter* is increasing over time, with an increasing public health concern (Moore et al., 2006). Research into the prevalence of resistance, the molecular mechanisms involved in resistance and their ability to spread is critical for the understanding and prevention of the spread of antimicrobial resistant *Campylobacter*.

Epidemiological surveillance of *Campylobacter* is necessary to detect and control outbreaks, and to monitor trends to identify changes in the antimicrobial resistance patterns. Several typing methods are being used for this purpose, and include pulsed-field gel electrophoresis, restriction fragment length polymorphism analysis, antigen gene sequence typing and multilocus sequence typing (MLST). In addition to traditional typing methods, whole genome sequencing of *Campylobacter* can increase our knowledge about their evolution and pathology, and give a better understanding of their ability to cause disease. With this information, the detection and control of the campylobacteriosis outbreaks can be more efficient.

1.2 Campylobacter

Campylobacter spp. are small (0.2 – 0.8 μm wide × 0.5 – 5 μm long), gram-negative, spirally shaped bacteria (Bolton, 2015), which are motile due to their uni- or bipolar flagellae (Figure 1) (Allos, 2001). They belong to proteobacteria in order of *Campylobacteriales*, along with *Helicobacter* and *Wolinella* (Young, Davis, & Dirita, 2007). They grow slowly and require microaerophilic conditions with a temperature between 37 and 42 °C (Davis & DiRita, 2008).

Campylobacter are present in the intestines of wild and domesticated animals including poultry, cattle, and swine, where they are part of the normal gut microflora (Cody, Bray, Jolley, McCarthy, & Maiden, 2017). Moreover, *Campylobacter* can be detected in the natural environments – they are found in surface waters as well as in sand from beaches (Kwan et al., 2008).

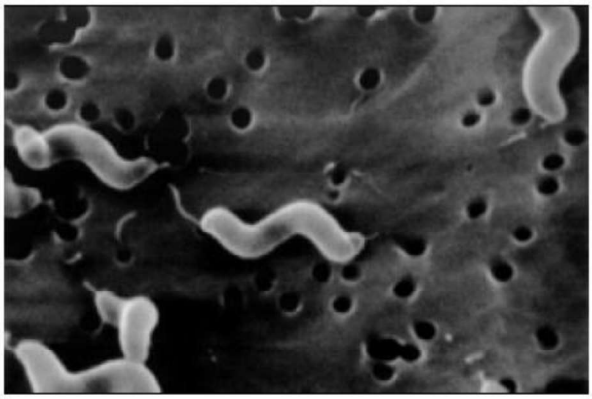

Figure 1. ***Campylobacter jejuni.*** Figure from (Altekruse, Stern, Fields, & Swerdlow, 1999).

1.2.1 Virulence and survival factors of *Campylobacter*

Campylobacter can survive in the gastrointestinal tract of poultry without causing illness (Meade et al., 2009), yet, in humans, they can lead to infections with doses as low as 500-800 bacteria (Robinson, 1981). The incubation period before the diarrhoeal symptoms in humans is 2-5 days. The differences in outcomes of infection with *Campylobacter* in man and chickens are not fully understood; however, it is believed that those variations might be associated with different bacterial gene expression in different hosts (Humphrey, O'Brien, & Madsen, 2007). Independently of the host, *Campylobacter* are able to colonise due to their virulence and survival mechanisms (Bolton, 2015).

Most *Campylobacter* are motile due to one or two polar flagella and their helical cell shape (Allos, 2001). Flagella allow the rotary cell movement while helical cell shape provides

corkscrew rotation (Ferrero & Lee, 1988). *Campylobacter* use chemotaxis to sense and move towards beneficial conditions. Motility, along with chemotaxis, are essential survival factors under the different chemotactic conditions in the gastrointestinal tract, and therefore, enable colonization of their host (Chang & Miller, 2006). Moreover, flagella are reported to be used by *Campylobacter* also as an export apparatus during secretion of specific proteins during host invasion (Poly & Guerry, 2008). Once the bacteria find their favourable conditions, they adhere to host gastrointestinal epithelial cells, employing several adhesins on their surface (Jin et al., 2001).

Campylobacter are known to produce the cytolethal distending toxin (CDT) which disrupts the host's cell mitosis, leads to their apoptosis and consequently causes infection in host's gastrointestinal tract (Ge, Schauer, & Fox, 2008; Pickett & Whitehouse, 1999). In some cases, infection with *Campylobacter* triggers the host's immune system that can lead to GBS or MFS (Bolton, 2015).

Iron is essential for successful colonization and survival of *Campylobacter* in the host, and *Campylobacter* have developed uptake mechanisms for iron as well as zinc (Davis, Kakuda, & DiRita, 2009; Hermans et al., 2011). In addition to iron and zinc uptake systems, *Campylobacter* have also developed specific efflux pumps, which enable them to maintain resistance to bile salts, heavy metals, and even antimicrobial drugs (Lin, Michel, & Zhang, 2002).

Because of exposure to many stresses within the food chain, *Campylobacter* have strict control of stress response. Within the slaughtering process, these bacteria are introduced to varying oxygen conditions, e.g., reactive oxygen species like superoxide anion (O_2^-) or hydrogen peroxide (H_2O_2). In these situations, *Campylobacter* respond by inducing or increasing the activity of the antioxidant defence in the form of enzymes such as glutathione, catalase, cytochrome c peroxidases, what results in prolonged aerobic adaptation (Jones, Sutcliffe, Rios, Fox, & Curry, 1993). Stress caused by different pH values or by starvation is overcome by entering the viable but non-culturable state, in which *Campylobacter* have a lower degree of metabolic activity and therefore, can survive critical conditions (Chaveerach, Ter Huurne, Lipman, & Van Knapen, 2003).

1.2.2 Genomic diversity of *Campylobacter*

C. jejuni has the ability to take up DNA from the environment either in the form of a plasmid or chromosomal DNA is a matter of concern (Boer et al., 2002; Wang & Taylor, 1990). Recombination between *Campylobacter* strains occurs relatively often which, in turn, leads to their genetic diversity. To date (4th September 2018), there are 1268 genome assemblies available for *C. jejuni*, of which 149 are complete genomes in the NCBI database (NCBI, 2018b). Besides chromosomal DNA, 56 plasmid assemblies have been annotated for *C. jejuni* species in the NCBI GenBank. Sizes of plasmids range from 1.82 kb (NZ_CM007887.1) to 119 kb (NZ_CP014743.1). For *C. coli*, there are 828 genome assemblies, and 22 of them are complete (NCBI, 2018a). Sizes of *C. coli* genomes are slightly bigger than *C. jejuni,* and the median size is 1.71405 Mb. Moreover, for this species, there are 41 annotated plasmids with sizes between 1.307 kb (NC_008049.1) and 180.543 kb (NZ_CP017026.1).

1.2.3 Campylobacter jejuni

C. jejuni are highly involved in human acute diarrheal disease. The illness in humans can occur after ingestion of even low inocula (500-800 bacteria) (Black et al., 1988), at any age, with a peak in children younger than one year in age and in young adults between 15 and 30 years old (Tauxe, Hargrett-Bean, Patton, & Wachsmuth, 1988). *C. jejuni* clonal complexes ST-21, ST-45, and ST-61 are the most frequent genotypes isolated from the dairy farming environment that cause human diseases (Kwan et al., 2008). Moreover, infection with *C. jejuni* is recognized as one of the most common causes of autoimmune disorders, namely, GBS and MFS, which lead to acute flaccid paralysis (McCarthy & Giesecke, 2001). Estimations in GBS acquisition reveal that 1 in 1000 infections leads to GBS (Allos, 1997).

1.2.4 Campylobacter coli

Because of the predominance of *C. jejuni* in human campylobacteriosis (Bae et al., 2005), most studies focus on this species, and relatively little is known about *C. coli.* However, it is reported that clinical symptoms of infection caused by *C. coli* are very similar to those, from a *C. jejuni* infection (Gillespie et al., 2002). Moreover, these infections affect adults in higher proportion than children, compared to *C. jejuni* infection (Gillespie et al., 2002; Kärenlampi, Rautelin, Schönberg-Norio, Paulin, & Hänninen, 2007). Although *C. coli* were first isolated from pigs suffering from infectious dysentery, they are not believed to be

pathogens of pigs (On, 2005). The primary sources of *C. coli* infection in humans is poultry, followed by swine, and sheep (Roux et al., 2013).

1.3 Epidemiology of *Campylobacter* species

1.3.1 Incidence and clinical features

Campylobacteriosis is one of the most prevalent bacterial infection worldwide. The occurrence of infections caused by *Campylobacter* is more frequent than those caused by *Salmonella, Shigella,* or Shiga toxin producing *Escherichia coli*- STEC (Blaser, Wells, Feldman, Pollard, & Allen, 1983) (Figure 2). According to the World Health Organization (WHO), every year campylobacteriosis affects ca. 550 million people (World Health Organization, 2013); roughly 200,000 of infections are acquired annually in the European Union (EFSA, 2018).

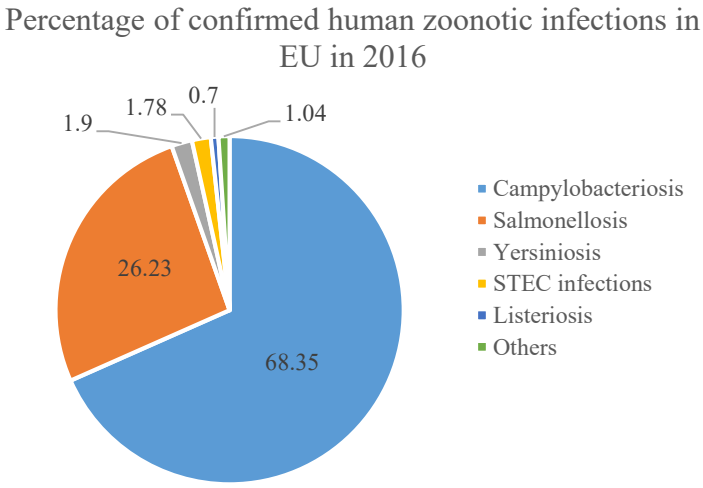

Figure 2. **Reported occurrences of infections caused by the most frequent zoonoses in European Union (EU) in 2016**. Figure based on data from EFSA & ECDC, 2017.

Recent studies on campylobacteriosis revealed seasonality, and it is believed that climatic variables influence the incidence of infection (Jore et al., 2010). In temperate European countries, regular infection incidences are observed in the summer months and reach up to 26,000 cases, while during winter the frequency of *Campylobacter* infections remains low and drops down to 12,000 cases (Figure 3). On the other hand, in tropical countries, minor infection variations are observed throughout the year (Strachan et al., 2013).

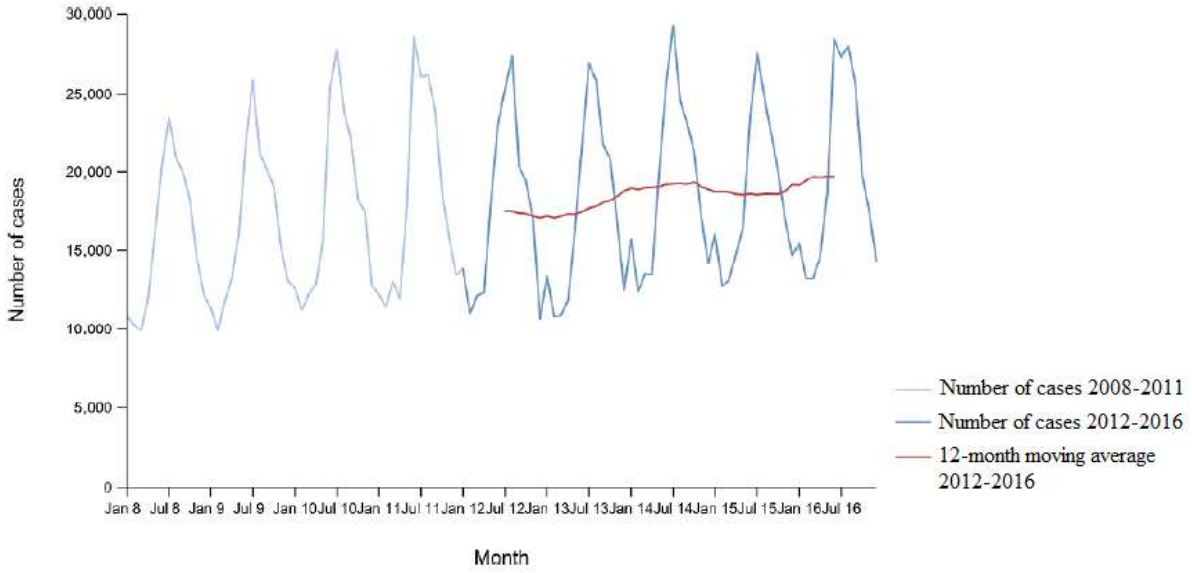

Figure 3. **Seasonality of campylobacteriosis in European countries between 2008-2016.** Figure from EFSA & ECDC, 2017.

Infections with *Campylobacter* differ in developed and developing countries. In former, campylobacteriosis usually manifests as self-limiting bloody diarrhoea affecting mainly young adults, while in the latter, it affects mainly children and manifests as watery diarrhoea (Blaser, 1997). The infection with *Campylobacter* usually lasts from 2.5 to 10 days.

1.3.2 Reservoirs

Significant disease outbreaks have been identified from retail meat, especially chicken and duck (Joensen et al., 2017), as well as from untreated drinking water (Kwan et al., 2008; MacDonald et al., 2015). It was discovered that *C. jejuni* present in the poultry gastrointestinal tract could enter the water supply and therefore infect humans in two ways: by consumption of contaminated poultry product or by drinking untreated water (Figure 4). Studies have reported the presence of *C. jejuni* in unpasteurised bovine milk, what can be another route of transmission to humans (Evans, Roberts, Ribeiro, Gardner, & Kembrey, 1996; Peterson, 2003). In Scandinavian countries and the United Kingdom, contamination with *Campylobacter* also occurs in raw red meat, with the dominance of beef, lamb, and pork (Kapperud, Skjerve, Bean, Ostroff, & Lassen, 1992; Smerdon, Adak, O'Brien, Gillespie, & Reacher, 2001).

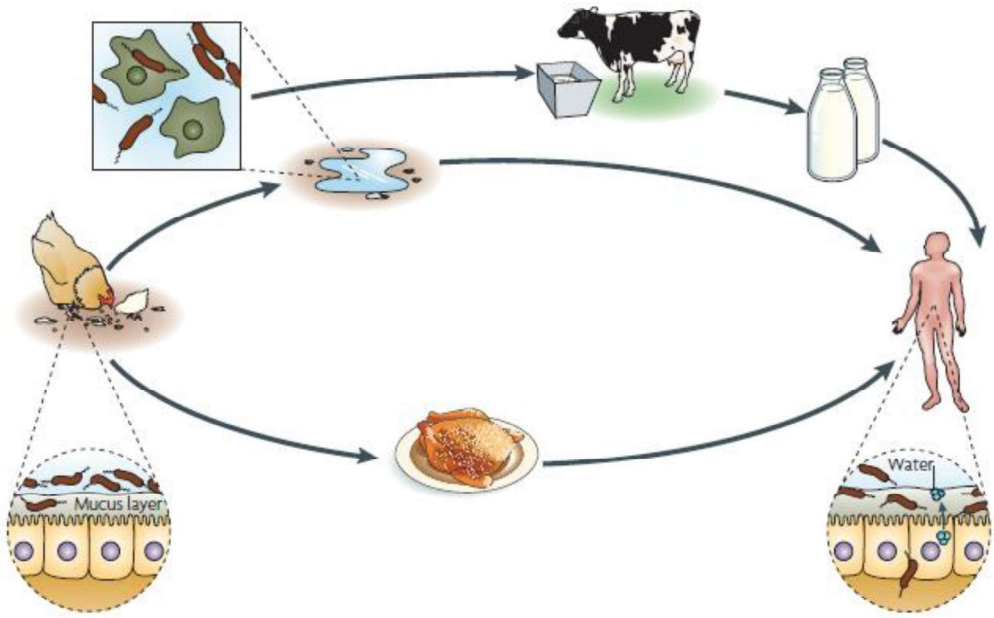

Figure 4. **The most frequent sources of campylobacteriosis caused by *C. jejuni***. Figure obtained from Young et al., 2007.

1.3.3 Epidemiology in Norway

The first records about campylobacteriosis in Norway are from 1979. Since then, the number of reported incidences to the Norwegian Surveillance System for Communicable Diseases (MSIS) has been increasing (Figure 5) (Norwegian Institute of Public Health, 2018). In 2017 it was reported 3884 incidences of campylobacteriosis in Norway, out of which 38% were domestically acquired, and 44% were travel-associated. The origin of the remaining 18% of cases was not known (Krosness et al., 2018). Since 2009, five major campylobacteriosis outbreaks were registered in Norway, and most of them were caused by *C. jejuni* (NIPH, 2009).

Similar to other European countries, in Norway, infections with campylobacteriosis have been associated with consumption of unpasteurised bovine milk, poultry, pork, and lamb, drinking untreated water, and direct contact with farm animals and dogs (Kapperud et al., 2003; Kapperud et al., 1992; Torp, Vigerust, Bergsjø, Er, & Hofshagen, 2014).

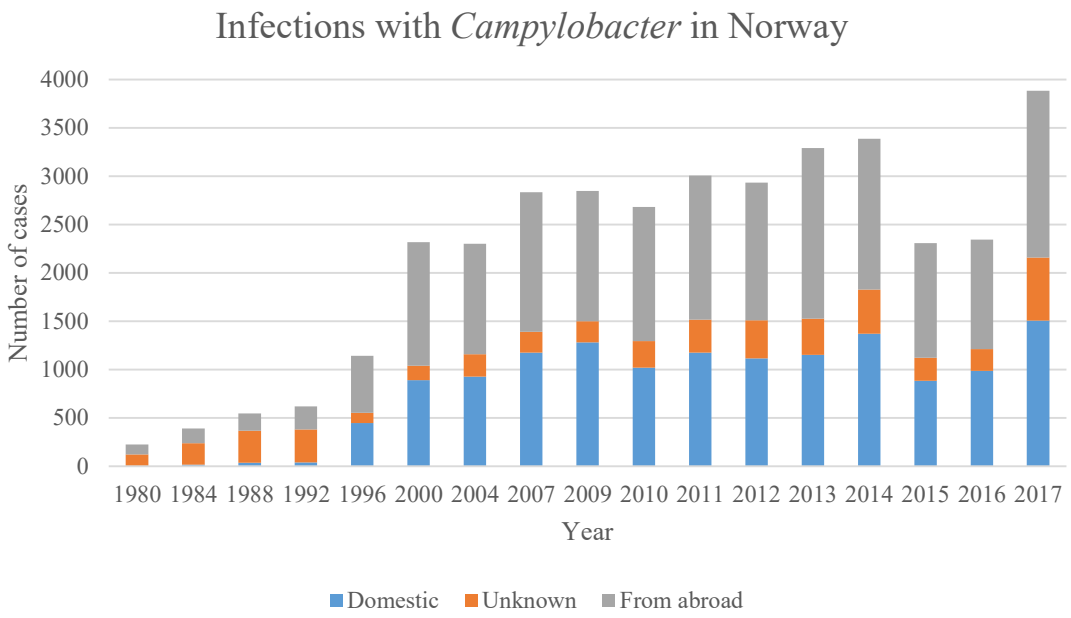

Figure 5. **The frequency of campylobacteriosis in Norwegian population between 1980 and 2017**. Figure based on data from NIPH, 2018.

In Norway, the occurrence of human campylobacteriosis rises gradually from spring, peaks in July and returns to the baseline level around the end of November (NIPH, 2018). The incidence of human infections in Norway seem to reflect the incidence of colonization in broiler flocks, what might suggest that seasons are factors influencing the infections. Similar data were obtained for other European countries like Denmark, Finland, Iceland, Sweden, United Kingdom and the Netherlands (Cody et al., 2012; Jore et al., 2010).

1.3.4 Antimicrobial resistance in human *Campylobacter* isolates

As reported by the European Centre for Disease Prevention and Control in 2016, the highest resistance of *C. jejuni* was observed for fluoroquinolones such as ciprofloxacin, tetracyclines, and macrolides including erythromycin (Table 1) (EFSA & ECDC, 2018). Although levels of the AMR varied between the European countries, the highest frequencies were observed in Portugal (94% for CIP and 82% for TET), Spain (84% for CIP and 78% for TET), and Italy (85% for CIP and 67% for TET) (EFSA & ECDC, 2018). Resistance towards ERM was relatively low (average 2.1%), however, in Norway, it reached 12% and was the highest noted value. Isolates of *C. coli* show a higher tendency to AMR compared to *C. jejuni*. In 2016 *C. coli* presented resistance to CIP and TET reaching even 100% in Portugal and Italy (EFSA & ECDC, 2018).

Table 1. **Antimicrobial resistance in *C. jejuni* and *C. coli* isolates based on data obtained in 2016 from 17 European countries**. Table based on information included in EFSA & ECDC, 2018.

Campylobacter	**Total percentage of isolates resistant to:**							
	N	GEN	N	CIP	N	ERM	N	TET
C. jejuni	6375	0.4	22676	54.6	21993	2.1	15614	42.8
C. coli	938	1.7	2565	63.8	2479	11.0	1919	64.8

N- a number of isolates tested; GEN- gentamycin; CIP- ciprofloxacin; ERM- erythromycin; TET- tetracycline.

A number of worldwide reports show that the AMR of *Campylobacter* species is increasing. In China between 1994 and 2010, the resistance of *Campylobacter* increased from 50% to 93% for CIP, and from 77% to 100% for TET (Zhou et al., 2016). In the United States, the resistance to CIP has raised from 13% to 16% between 2004 and 2012 (Geissler et al., 2017). In South Africa between 2002 and 2007 resistance to CIP raised from 8% to 13%, and to ERM from 25% to 53% (Bester & Essack, 2008). The resistance among *Campylobacter* in Europe is increasing as well. Reports from between 2013 and 2016 show relevant growths in *C. jejuni* resistance to CIP (from 55% to 95% in Estonia), and TET (from 25% to 45% in Austria and from 25% to 60% in Estonia) (EFSA & ECDC, 2018).

Antimicrobial resistance (AMR) within *Campylobacter* strains occurs probably due to the usage of antibiotic drugs (mostly quinolones) in veterinary medicine and the agricultural industry (Aarestrup & Engberg, 2001). AMR in *Campylobacter* is prevalent to a greater extent within countries approving quinolones in poultry production, compared to countries which do not use these antimicrobials (Alfredson & Korolik, 2007; Wieczorek, Kania, & Osek, 2013). Antimicrobial-resistant *Campylobacter* from animals can be transmitted to humans, leading to adverse consequences in therapy for human systemic infections (Bae et al., 2005).

1.4 Next-generation sequencing and genome assembly

Sanger's successful sequencing of bacteriophage ΦX174 by chain termination (Sanger, Nicklen, & Coulson, 1977) was the beginning of the sequencing era. In less than twenty years from that event, improvements in the sequencing technology led to the entire sequence of the first bacterial organism (Fleischmann et al., 1995) and have kept on developing ever since. By 2005, the time of sequencing was reduced from months or even years to hours or days decreasing the price about thousands of times comparing to Sanger's method (Loman et al., 2012). With the subsequent rise of high-throughput sequencing, plenty of next-generation

sequencing (NGS) platforms have appeared along with bioinformatic tools appropriate for them (Loman et al., 2012).

NGS technologies involve minor costs, shorter processing time and sequencing on the larger scale, compared to Sanger's sequencing. NGS can produce up to one billion of short reads in a single run, at a relatively affordable per-base cost (Metzker, 2010). It was estimated that obtaining the entire sequence of the human genome by 2003, reached a total cost of ~$450 million. With the evolution of sequencing technologies, those values kept decreasing, through $20-25 million in 2006, down to $1000 in 2016 (National Human Genome Research Institute, 2016).

The main drawbacks of NGS technology were higher error rates (0.1 – 15%) and shorter read lengths (35 – 700 bp) in comparison to Sanger's technology (Goodwin, McPherson, & McCombie, 2016). As a result, genome assembly was more complicated and required the development of new alignment algorithms (McPherson, 2009; Van Dijk, Auger, Jaszczyszyn, & Thermes, 2014).

Commercially available NGS technologies include, among others, pyrosequencing by Roche/454 FLX (Margulies et al., 2005), sequencing by synthesis by Illumina/Solexa (Quail et al., 2012), Sequencing by Oligo Ligation Detection (SOLiD™) by Applied Biosystems (Valouev et al., 2008) and semiconductor sequencing (Ion Torrent) by Life Technologies (Metzker, 2010). The standard methodology among NGS techniques involves i) preparation of NGS libraries in a cell-free system, ii) production of thousands or even millions of sequencing reactions in parallel and iii) ability to detect the output without the use of electrophoresis (Van Dijk et al., 2014).

Innovations in NGS technologies led to the growth of single molecule sequencing known as third-generation sequencing. These methods generate greater read lengths (several thousand bp) without former amplification of the DNA (Land et al., 2015; Pareek, Smoczynski, & Tretyn, 2011). Although still evolving, third generation sequencing and its methodology are already described with an example of Helicos Heliscope™ by Helicos BioSciences, MinION by Oxford Nanopore Technologies and PacBio by Pacific Biosciences (Pareek et al., 2011; Pushkarev, Neff, & Quake, 2009; Schadt, Turner, & Kasarskis, 2010).

1.4.1 Whole genome sequencing

Evolution of NGS technologies led to the development of the whole genome sequencing (WGS) techniques. WGS data gives insights not only into functions and evolution of bacteria but also into their interactions with host, environment and each other. Moreover, with more efficient and affordable techniques, multiple genomes of the same genus are being collected, allowing to evaluate horizontal gene transfer within one or more species (Loman & Pallen, 2015). This knowledge triggered the use of WGS methods in bacterial surveillance and outbreak detection and investigation, as they allow to identify bacterial transmission pathways more accurately (Harris et al., 2013).

1.4.2 Illumina Sequencing

Illumina is one of the most widely-used NGS platforms on the biological market. This technique employs the sequencing-by-synthesis approach which gives higher accuracy and fewer error calls (Illumina Inc., 2010; Mardis, 2008). Illumina technique involves clonally amplified DNA templates which are immobilized on the surface of a glass flow cell, which enables a bridge amplification. Subsequently, fluorescently labelled reversible-terminator nucleotides are added simultaneously along with the DNA polymerase to join with the templates on the flow cell. The fluorescence signal measurements are made after each incorporation of the single nucleotide, and eventually, the blocking group of the nucleotide is removed before the next incorporation. The following steps are proceeded for a specific number of cycles until reads of 150-250 bases length are obtained (Illumina Inc., 2010; Mardis, 2008; Quail et al., 2012). The simplified process of Illumina sequencing is presented in Figure 6.

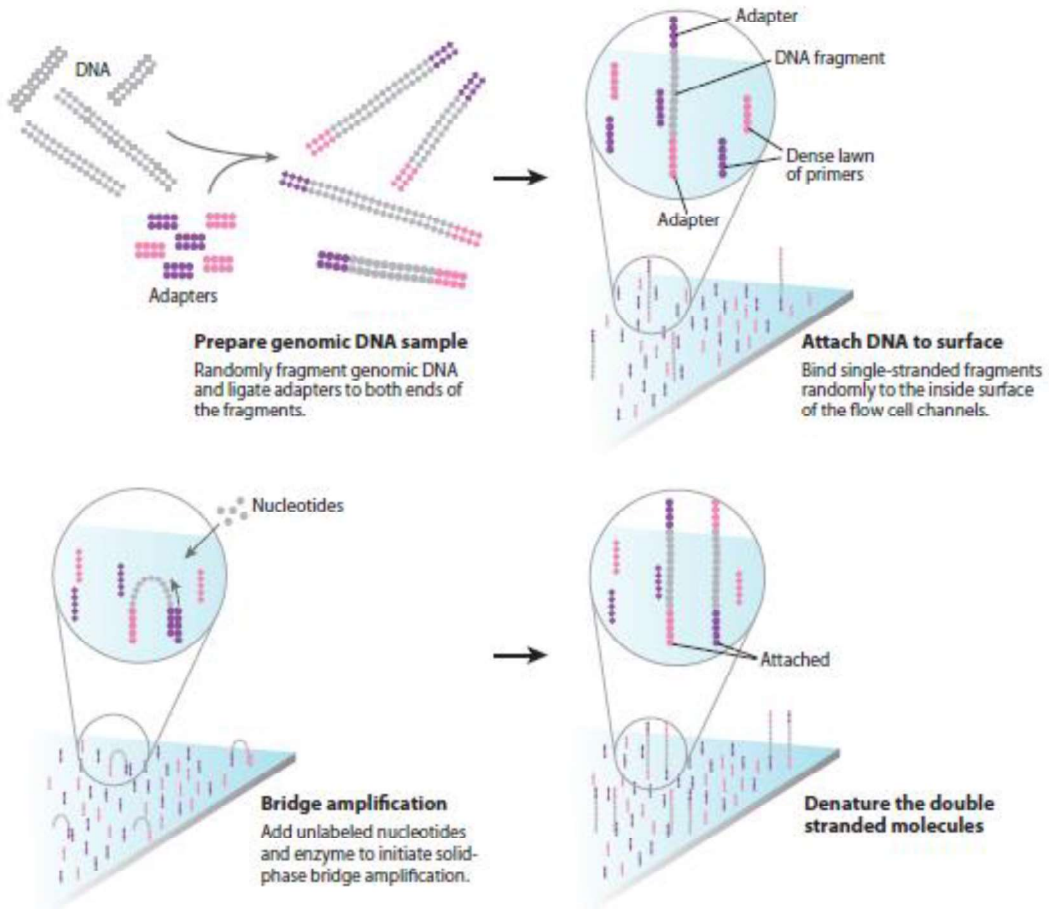

Figure 6. **Simplified scheme of the Illumina sequencing approach**. The template DNA is first randomly fragmented, and specific adapters are ligated to both ends of the DNA fragments. Subsequently, these fragments are attached to the surface of the flow cell, and four fluorescently labelled, 3'-OH blocked nucleotides are added along with the DNA polymerase. One specific nucleotide incorporates with the template and unbound nucleotides as well as polymerase are washed away. The fluorescence signal is then measured, and the 3' blocking group is removed before following nucleotide incorporation. Figure from Mardis, 2008.

1.4.3 Genome assembly

Sequencing gives output in the form of reads, the length of which varies depending on the sequencing platform (Table 2). Regarding the sequencing objectives, there are two classes of reads, namely, single-end and pair-end reads. The difference between them is that single-end reads refer to one sequence end of the DNA fragment, while pair-end reads refer to both sequence ends (He, Zhang, Peng, Wu, & Wang, 2013). Moreover, there are also mate-pair reads, which similarly to pair-end reads, introduce information from both ends of the read, but they are much longer – up to several thousands of base pairs (Ekblom & Wolf, 2014).

Generally, assembling the genome is performed by grouping the reads into overlapping DNA fragments called contigs, and subsequently grouping the contigs into larger but discontinuous DNA fragments called scaffolds, and finally putting scaffolds into a complete

DNA sequence (Figure 7) (He et al., 2013). Most of the genome assemblers give a set of scaffolds in FASTA format as a final output, where Ns represent gaps between contigs.

NGS methods generate millions of short reads, assembly of which is extremely laborious. One of the most common complications concerns the repetitive regions in the genome, the assembly of which seems to be impossible with the use of NGS short reads. Therefore, genomes of high quality, deprived of any gaps constitute only 35% of all sequenced genomes (Koren & Phillippy, 2015).

Table 2. **Overview of sequencing platforms and their output.** Table based on data from Buermans & Den Dunnen, 2014; Goodwin et al., 2016; Mardis, 2011; Wajid, Sohail, Ekti, & Serpedin, 2016.

Producer	Sequencing platform	Run types	Run time	Read length (bp)
Roche	454 FLX	Single-end	~8.4 hours	250 – 330
Illumina/Solexa	HiSeq2000/2500	Single & Pair-end	12 days	2 × 100
	MiSeq	Single & Pair-end	65 hours	2 × 300
Applied Biosystems	SOLiD	Single & Pair-end	14 days	50
Life Technologies	Ion Torrent	Single-end	4 hours	200 – 400
Helicos	HeliScope	Single-end	10 days	~30
Pacific Biosciences	PacBio RSII	Single-molecule real-time reads	3 hours	15 kb
Oxford Nanopore	MinION	Single-molecule real-time reads	~48 hours	Up to 200 kb

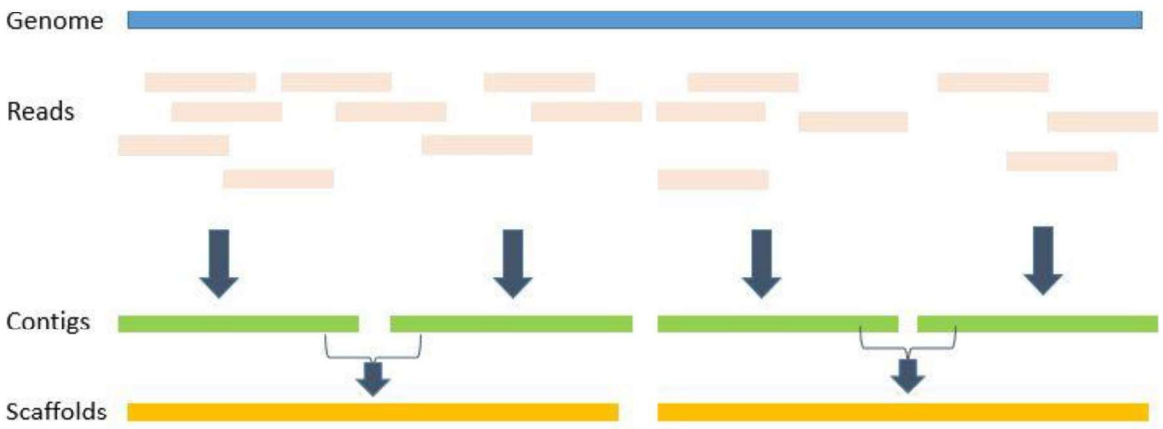

Figure 7. **Simplified process of assembling the genome**. Randomly fragmented genomic DNA is sequenced, resulting in an abundant number of short reads. The *de novo* assembly algorithms merge those reads into longer fragments- contigs, which are subsequently joined into scaffolds. Ideally, all joined scaffolds will result in a completed sequence of genome DNA.

1.4.3.1 *De novo* assembly

The NGS techniques generate millions of short reads, which make the assembly process complicated and requiring well-adapted algorithms and statistical tools (Wajid et al., 2016). Modern sequence assemblers are based on paradigms, the choice of which is dependent on features of reads being assembled. The most recognized paradigms are greedy, overlap-layout-consensus (OLC), *de Bruijn* graph (DBG) and string graph (Nagarajan & Pop, 2013). Assemblers engaging greedy, consider only local links between reads and join only the most overlapping reads if they do not disrupt already joined reads (Nagarajan & Pop, 2013). OLC assemblers (Figure 8A) find all overlapping reads, lay them into a graph and finally generate the consensus sequence basing on Multiple Sequence Alignment. The graph construction enables to take into account the global relationship between the reads (He et al., 2013). The simplified version of the OLC paradigm (string graph) removes redundant information (transitive edges) from the global overlap graph (Nagarajan & Pop, 2013). Finally, assemblers based on DBG (Figure 8B) divide the reads into smaller *k*-mers, which are subsequently involved in constructing the graph. Two adjacent *k*-mers differ by *k*-1 bp and are laid out along the genome. The graph constructed this way is used as a base to build the whole genome sequence (He et al., 2013).

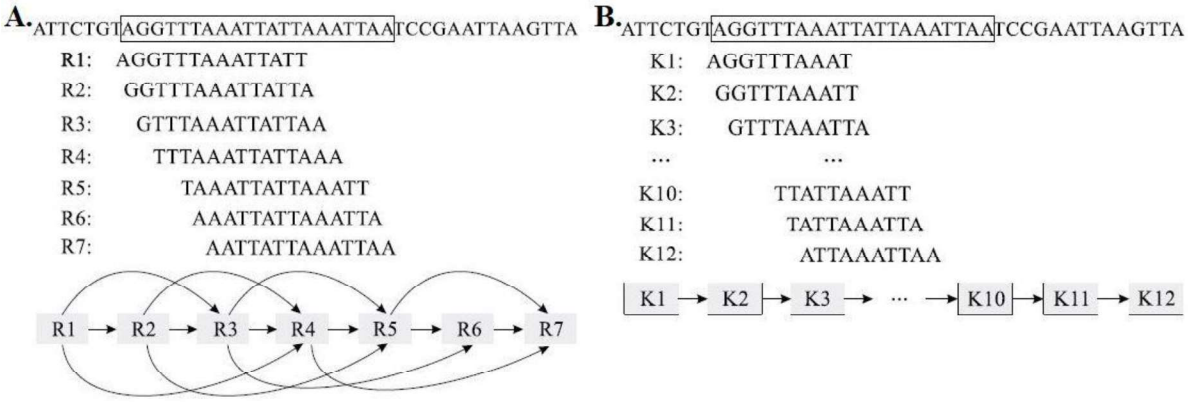

Figure 8. **The visual presentation of two paradigms used by *de novo* assemblers**. **A.** Overlap-layout-consensus (OLC) paradigm finds all overlapping reads (R1-R7), lists them along the genome and finally deduces consensus sequence. Curved arrows refer to transitive edges (overlaps that are covered by a set of shorter overlaps). **B.** *de Bruijn* graph divides the reads into shorter *k*-mers (K1-K12) and lists them along the genome by distinguishing their adjacent relations. For example, *k*-mers K1 and K2 share exactly 9 bp, therefore two adjacent *k*-mers will differ by *k*-1 bp. Note: the principles of paradigms here are shown with the use of a 21 bp fragment of the genome, 14 bp reads and 10 bp *k*-mers. Figure obtained from He et al., 2013.

SPAdes is a command line *de novo* assembler that bases on DBG, and applies to both, standard bacterial and single cell sequencing datasets (Bankevich et al., 2012). SPAdes bears paired-end reads, mate-pairs and unpaired reads in FASTA/FASTQ format, and is compatible

with Illumina, IonTorrent, Oxford Nanopore or Sanger reads. SPAdes v 3.9.0. comes in septate modules, namely BayersHammer, IonHammer, SPAdes and MismatchCorrector, where "hammer" modules are used to obtain high-quality assemblies with read error correction.

1.4.3.2 Reference-based mapping

Reference-based mapping involves the alignment of the reads to a reference genome and subsequent incorporation of those reads into final contigs (Li, Ruan, & Durbin, 2008). Although this method is less complicated than *de novo* assembly, it still requires well-designed alignment programs because of challenges such as a large number of short reads produced by NGS technologies. The algorithms used in reference-based mapping programs have to be highly efficient and accurate to cope with repetitive fragments of the genome. A reliable alignment can be obtained by avoiding the mapping of those repetitive reads to multiple positions in the genome (Koren & Phillippy, 2015). What is more, the accuracy of reference-based mapping methods strongly depends on the availability and quality of the reference genome as well as on the quality of the reads (Benjamin, Nichols, Burke, Ginsburg, & Lucas, 2014). A single mutation or sequencing error can result in mapping the read into the wrong position of the alignment (Koren & Phillippy, 2015).

Most of the algorithms for reference-based mapping find the potential mapping location by exact matching or scoring the sequence similarity with employment of the sequence of each read (Benjamin et al., 2014). The reference-based mapping methods are well reviewed by Benjamin et al. and they can be divided among others, into unspliced aligning. This kind of mapping includes the Burrows-Wheeler Transform Method that aligns the reads to a reference without permitting large gaps, with former indexing of the reference sequence. Subsequently, the algorithm looks for perfect matches to merge the reads into the final contigs (Benjamin et al., 2014). An exemplary assembler based on Burrows-Wheeler Transform Method is a Burrows-Wheeler Aligner (BWA) (Li & Durbin, 2009). BWA is a software based on Smith-Waterman local alignment containing three variants: BWA-backtrack (designed for Illumina reads up to 100 bp), BWA-MEM and BWA-SW (both, for reads between 70 bp and 1 Mbp). BWA-MEM is usually preferred algorithm because of its accuracy and speed (Li & Durbin, 2009).

1.5 Molecular typing

Molecular typing is widely used in surveillance by public health institutions, as it provides essential information about bacterial isolates in terms of local and global epidemiology (Maiden et al., 1998). Typing methods allow the investigation of phylogeny, population genetics, and their spread, and allow for the differentiation between isolates from different common or unlinked sources, (Unemo & Dillon, 2011). Furthermore, the knowledge obtained by means of molecular typing can give an overview of disease outbreaks, virulence factors, and antimicrobial susceptibility.

Molecular typing methods are based on identifying polymorphisms in a single locus or multiple loci of the genome and can be divided into i) DNA-fingerprint based typing methods and ii) DNA sequence-based methods.

Multilocus sequence typing (MLST) is a DNA sequence-based typing method that examines the genome using multiple housekeeping loci (Maiden et al., 1998). This methodology enables to trace and analyze the patterns of genetic exchange between bacteria in the way that is standardized, reproducible and portable (Belén, Pavón, & Maiden, 2009). Data obtained by MLST can be stored and shared by online databases such as PubMLST, enabling an overview of the global epidemiology of specific species (Adzitey, Huda, & Ali, 2013; Jolley, Chan, & Maiden, 2004; Maiden et al., 2013).

The basic MLST scheme is built on the identification of allelic profiles of seven different loci, for which a unique number is assigned. Subsequently, those unique numbers of seven loci are involved into creation of the allelic profile (for example, 2-3-4-3-8-4-6), or a sequence type (ST) with a numerical specification (for example, ST11) (Maiden et al., 2013). The relationship between typed isolates is usually presented as a dendrogram, which is a result of the pairwise comparison of allelic profiles. This is a convenient way of distinguishing isolates, the profiles of which are very similar (Enright & Spratt, 1999).

1.5.1 Core genome multilocus sequence typing

The core genome MLST (cgMLST) is a development of the traditional MLST typing and is based on WGS data (Maiden et al., 2013). This technique delivers differentiation of the strains and isolates of the same species and relies on a gene-by-gene comparison of allelic profiles, based on a fixed number of conserved chromosomal genes. In this method, a core genome is defined as a set of coding loci present in a majority of analysed bacterial isolates (Mellmann et al., 2016).

A recent cgMLST typing scheme for *C. jejuni* and *C. coli* has generated a set of 1343 genes being found in 95% of 2472 human *Campylobacter* isolates (Cody et al., 2017). The core genome established by these authors was based on re-annotated *C. jejuni* NCTC 11168 reference genome and consisted of 96.9% of genes with putative functions, such as metabolism of amino acids, their derivatives, proteins, cofactors, and vitamins (Cody et al., 2017).

1.6 Phylogeny

Phylogenetic analysis is essential for the investigating evolutionary relationships of species and is widely used in comparative genomics, functional prediction, detection of lateral gene transfer and other biological research (Bear et al., 2016; Dereeper et al., 2008). Phylogenetic approaches involve identification of homologous sequences, subsequent multiple alignment of those sequences, hypothetic phylogenetic reconstruction and graphical representation in the form of a phylogenetic tree (Figure 9) (Dereeper et al., 2008).

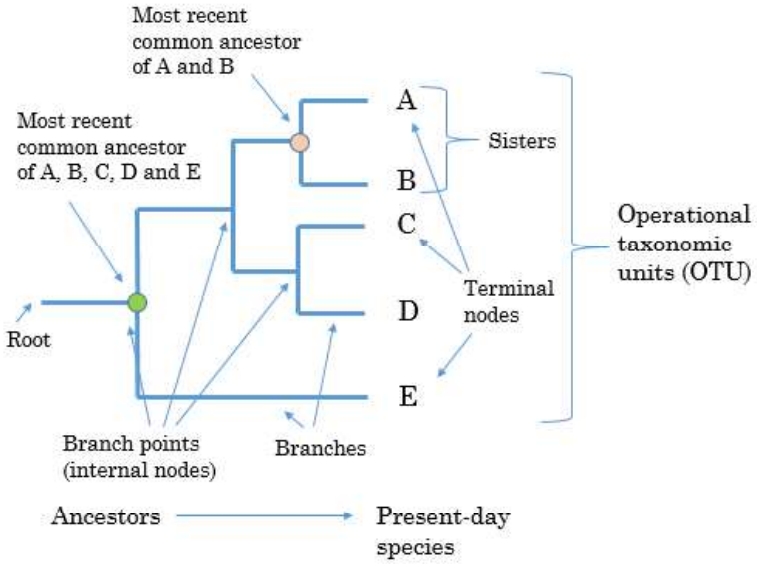

Figure 9. **Representation of a simple rooted and unscaled phylogenetic tree**. The tree above represents evolutionary relationships between operational taxonomic units (OTU) (A – E) placed on the tips of branches. OTU can relate to genes, proteins, organisms or species. Each internal node represents an event of divergence of one group into two descendant groups. The top node marked in pink shows the most recent common ancestor for taxa A and B (which therefore become the closest relatives – sisters), while the node marked in green represents the most recent common ancestor for all analysed OTU A – E.

A phylogenetic tree is a way to visualize the relationships between taxa, their evolutionary distance or even their common ancestor. Rooted phylogenetic trees enable to distinguish the common ancestor of homologous taxa, while unrooted trees give an overview

on the degree of the evolutionary relationship regardless of the direction of the evolutionary timeline (Graham, Olmstead, & Barrett, 2002).

1.6.1 Construction of phylogenetic trees

Phylogenetic trees can be estimated with the use of various methods, the choice of which is dependent on the character of the study. The major phylogeny algorithms are as follows: maximum parsimony (MP), maximum likelihood (ML), neighbour-joining (NJ) and unweighted pair group method using arithmetic average (UPGMA) (Kuhner & Felsenstein, 1994; Xiong, 2006). The UPGMA trees are built by continuous clustering through stepwise reduced distance matrices. This technique assumes that all taxa evolve at a constant rate and is the simplest clustering method (Xiong, 2006). The MP approach is based on an evaluation of all possible topologies and choice of the tree with the fewest evolutionary changes or the shortest overall branch lengths, calculated from the character matrix (Saitou & Nei, 1987). The NJ trees are created using the principle of minimum evolution, which is similar to MP in that way that it also chooses the optimum tree basing on the minimum overall branch lengths, however, it estimates trees from a distance matrix (Xiong, 2006). The ML method, on the other hand, tries to choose the best tree with the highest likelihood utilizing probability models. This approach employs substitution models and is believed to be the most reliable phylogenetic technique (Huelsenbeck & Rannala, 1997).

Statistical evaluation of phylogenetic tree construction is relevant in assessing its reliability. This assessment can be established by proceeding bootstrapping, which is a statistical method testing for sampling errors in a phylogenetic tree. Bootstrapping is based on reconstructing multiple trees (replicates) from variations of the input data (Pattengale, Alipour, Bininda-Emonds, Moret, & Stamatakis, 2010). Despite these variations, the strong phylogenetic relationships should be supported and unchanged; weak relationships between taxa will result in the creation of different trees, compared to the original topology (Xiong, 2006).

1.7 The aim of the study

The aim of this study is to describe the population structure of a selection of *C. jejuni* and *C. coli* isolates from Norway isolated between 2009 and 2017, using whole genome sequencing in order to identify cluster thresholds for outbreak detection and genetic mechanisms of antimicrobial resistance.

The aim will be fulfilled by completing the following objectives:

- To analyse assembly of sequencing reads utilizing *de novo* assembly and reference-based mapping.
- To implement a core genome multilocus sequence typing (cgMLST) scheme for *Campylobacter* at the Reference Laboratory for Enteropathogenic Bacteria at the Norwegian Institute of Public Health.
- To compare the phylogeny of *Campylobacter* isolates between different clustering models (MP, ML, NJ, and UPGMA).
- To identify *in silico* antimicrobial resistance genotypes using BLAST algorithms against ResFinder and CARD online databases, and compare results with phenotypic resistance for a selection of antibiotic profiles based on clinical resistance and epidemiological cut-offs.

2. Material and Methods

A basic outline for methods performed in this study are depicted in Figure 10.

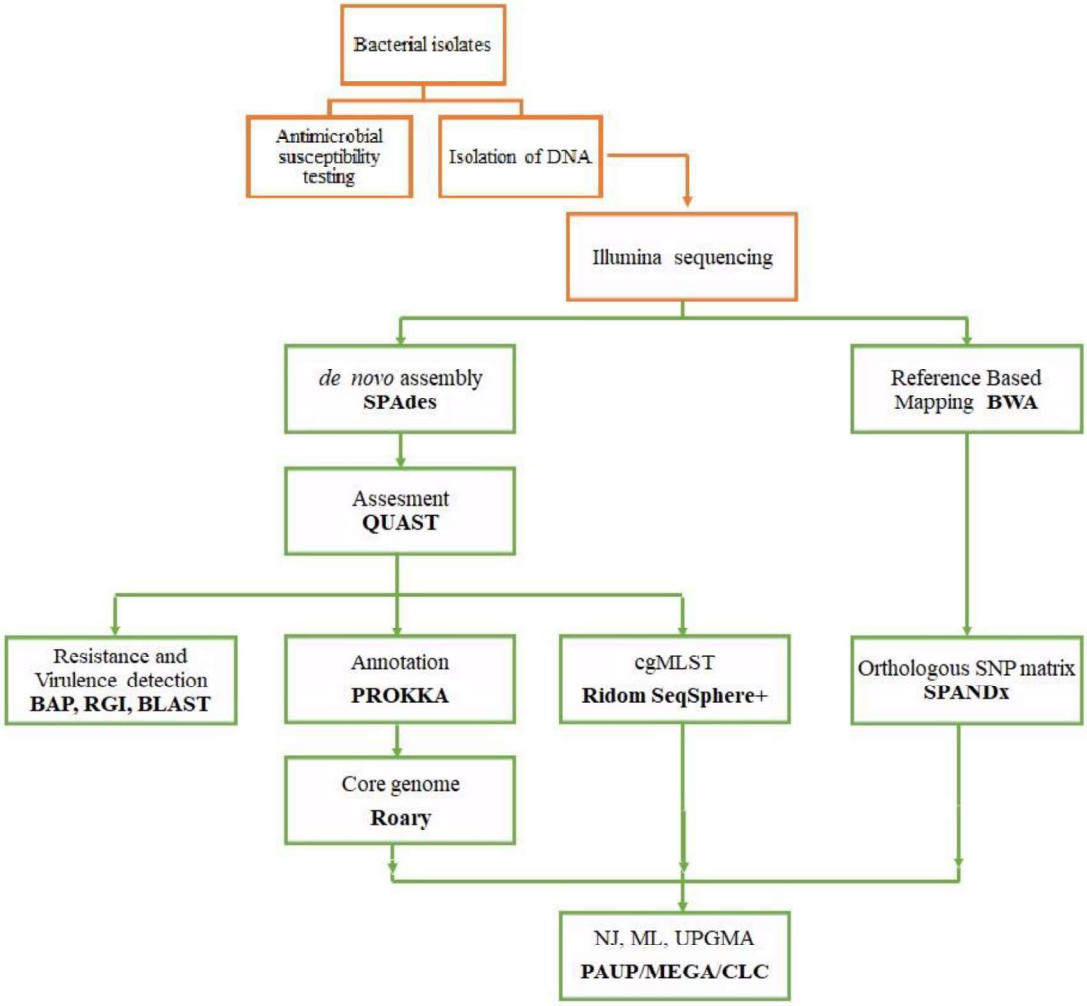

Figure 10. **A general outline of methods performed during the study.** Wet lab methods and sequencing by Illumina are marked in orange, while subsequent *in silico* methods are marked in green. Paired-end reads obtained from sequencing were assembled both, *de novo* and mapped to the reference. Phylogenetic relationships were based on core genome alignment and orthologous SNP matrix.

2.1 Bacterial isolates

For this study, we included 40 clinical *Campylobacter* isolates collected from patients across Norway sent to NIPH for surveillance purposes. 35 *C. jejuni* and five *C. coli* were selected based upon their antimicrobial resistance profiles, and diversity with regards to sporadic isolates and isolates collected from previous outbreaks, and isolates acquired in Norway and acquired abroad. One isolate was included from a non-human source. An overview of all isolates is presented in Table 3.

Table 3. **Description of included isolates (n=40) of *C. jejuni* and *C. coli* in this study.**

Isolate ID	Pathogen	Year/month of collection	Source	Outbreak	Country of acquiring the infection
1109-1129	*C. jejuni*	2009/05	Human	Kindergarten	Norway
1109-1130	*C. jejuni*	2009/05	Human	Kindergarten	Norway
1109-1179	*C. jejuni*	2009/05	Human	Kindergarten	Norway
1109-1180	*C. jejuni*	2009/05	Human	Kindergarten	Norway
1109-1207	*C. jejuni*	2009/06	Non-human	Kindergarten	Norway
12EP000401	*C. jejuni*	2012/08	Human		Italy
12EP000408	*C. jejuni*	2012/08	Human		Turkey
12EP001377	*C. coli*	2012/11	Human		India
13EP000100	*C. jejuni*	2013/01	Human		-
13EP000133	*C. jejuni*	2013/01	Human		India
13EP001161	*C. jejuni*	2013/07	Human		Norway
13EP001259	*C. jejuni*	2013/07	Human		Canada
13EP001978	*C. jejuni*	2013/09	Human		Ecuador
13EP002420	*C. jejuni*	2013/10	Human	Restaurant	Norway
13EP002423	*C. jejuni*	2013/10	Human	Restaurant	Norway
13EP002426	*C. jejuni*	2013/10	Human	Restaurant	Norway
13EP002526	*C. jejuni*	2013/10	Human	Restaurant	Norway
13EP002546	*C. jejuni*	2013/11	Human	Restaurant	Norway
14EP000043	*C. coli*	2014/01	Human		Tanzania
14EP000843	*C. jejuni*	2014/04	Human		Spain
14EP001612	*C. jejuni*	2014/08	Human		Poland
14EP001617	*C. jejuni*	2014/08	Human		Norway
14EP001642	*C. jejuni*	2014/08	Human		Norway
15EP000113	*C. jejuni*	2015/01	Human		Spain
15EP000253	*C. jejuni*	2015/02	Human		Philippines
15EP000596	*C. coli*	2015/04	Human		Cuba
15EP001566	*C. jejuni*	2015/09	Human		Norway
15EP002192	*C. jejuni*	2015/12	Human		Germany
16EP000145	*C. jejuni*	2016/01	Human		Norway
16EP000265	*C. jejuni*	2016/02	Human		Turkey
16EP000713	*C. coli*	2016/05	Human		Spain
16EP001088	*C. jejuni*	2016/07	Human		Norway
16EP001139	*C. jejuni*	2016/07	Human		Norway
16EP001848	*C. jejuni*	2016/10	Human		Norway
16EP001980	*C. coli*	2016/10	Human		Thailand
16EP002233	*C. jejuni*	2016/12	Human		Malawi
17EP001087	*C. jejuni*	2017/06	Human	Steinkjer	Norway
17EP001093	*C. jejuni*	2017/06	Human	Steinkjer	Norway
17EP001096	*C. jejuni*	2017/06	Human	Steinkjer	Norway
17EP001113	*C. jejuni*	2017/06	Human	Steinkjer	Norway

2.2 Isolation of DNA

Isolation of the DNA and antimicrobial susceptibility testing was performed at NIPH, with the supervision of enteropathogenic bacteria department personnel. Isolates (n=40) were cultured on non-selective blood agars at 42℃, overnight under microaerophilic conditions and checked for contamination prior to re-culturing on fresh media in the same conditions.

Subsequently, isolated colonies were dissolved in bacteria lysis buffer (150 µl), Tris EDTA buffer (150 µl), and Proteinase K (30 µl) and incubated for 10 minutes at 60°C and 105°C before freezing at -80 °C. Further steps of DNA isolation were proceeded according to the MagNA Pure 96 System (Roche) manual. Supplementary file S1 includes DNA concentrations for each isolate, before library preparation and sequencing.

2.3 Antimicrobial susceptibility testing

For antimicrobial susceptibility testing, isolates cultured on non-selective media at 42°C overnight were re-cultured on Mueller-Hinton agars with 5% defibrinated horse blood and 20 mg/l β-NAD. The inoculum was suspended in saline to 1.0 McFarland standard. Further procedures were following the manual of the European Society of Clinical Microbiology and Infectious Diseases (EUCAST) disk diffusion method v. 6.0 (EUCAST, 2017), with the use of antibiotic strips. The minimum inhibitory concentrations (MICs) were estimated for the following antimicrobial agents: ciprofloxacin, tetracycline, erythromycin, gentamicin and nalidixic acid (NX). These antibiotics were chosen because of their use as therapeutic drugs for campylobacteriosis cases and because of recent bacterial AMR towards these antibiotics. Epidemiological cut-offs were determined, based on EUCAST standards. *C. jejuni* ATCC 33560 was used as the quality control organism.

2.4 Whole genome sequencing

Library preparation and sequencing of genomic DNA of 40 *Campylobacter* isolates were performed by personnel of NIPH using KAPA Library Preparation Kit (KAPABiosystems, 2017), and the Illumina MiSeq sequencing platform following the manufacturers' manuals. The output was acquired as pair ended libraries with pair-end reads of 150 and 250 bases long, respectively. These reads were subsequently used for genome assembly, determination of cgMLST and construction of phylogenetic trees.

2.5 Assembly of *Campylobacter* genome

2.5.1 Quality control and improvement

Pair-ended reads acquired from NIPH were examined for quality using FastQC. To remove remaining sequences of Illumina's adapters and to improve the quality of the reads, software Trimmomatic v 0.32 (Bolger, Lohse, & Usadel, 2014) was used with the command line as follows:

~java –jar Trimmomatic-0.32.jar PE [path to forward reads] [path to reverse reads] [directory for output reads] ILLUMINACLIP:TruSeq2-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:3:15 MINLEN:36.

Trimmed sequences were subsequently re-assessed for their quality by FastQC before further analysis.

2.5.2 De novo assembly

The reads were assembled with SPAdes v 3.9.0., with a command as follows: **~spades.py --k 21, 33, 55, 77, 99, 127 --careful -1 [path to forward reads] -2 [path to reverse reads] --cov-cutoff 5 –o [path to output file]**

Additional pipeline options were chosen to i) reduce the number of mismatches and short indels in final contigs (*--careful*), ii) increase *k*-mers lengths depending on the coverage (*-k 21, 33, 55, 77, 99, 127*), iii) reduce potentially misassembled low coverage (threshold 5) contigs (*--cov-cutoff 5*). To assemble plasmids from WGS reads, **~plasmidspades.py** was used in a similar manner. These predicted plasmid sequences were used as queries in BLASTn (Altschul, Gish, Miller, Myers, & Lipman, 1990) searches against a publicly available database (NCBI, 2018c) to find homologous plasmids sequences. The threshold values from BLASTn were set to ≥ 80% coverage and ≥ 95% identity. The plasmid of significance was defined by size of ≥ 20.0 kbp.

A5-miseq pipeline (Coil, Jospin, & Darling, 2014) was proceeded for randomly chosen isolates: 1109-1130, 13EP000100, 13EP001161, 13EP002526 and 14EP001617, to compare this method with *de novo* assembly by SPAdes.

For assessment of assemblies performed by SPAdes and A5-miseq, QUAST v 4.3. (Gurevich, Saveliev, Vyahhi, & Tesler, 2013) was used with the following command:

~quast.py [path to the contig file] -R [path to the reference genome file] -m 1000

In order to find the most accurate reference genomes for assessment purposes, during QUAST analysis different *Campylobacter* genome sequences were used, namely: *C. jejuni* NCTC 11168 (NC_002163.1), *C. jejuni* TS1218 (NZ_CP017860.1), *C. jejuni* RM1221 (NC_003912.7) and *C. jejuni* 81-176 (NC_008787.1). For *Campylobacter coli*, on the other hand, following genome sequences were used: *C. coli* ASM202418v1 (NZ_CP019977.1), *C. coli* CVM N29710 (NC_022347.1), *C. coli* FB1 (NZ_CP011015.1) and *C. coli* RM5611 (NZ_CP007179.1).

2.5.3 Reference-based mapping

BWA-MEM algorithm was used to map Illumina paired-end reads to *C. jejuni* NCTC 11168 and *C. coli* CVM N29710 as reference genomes. Mapping was performed with default parameters utilizing the following commands:

~bwa index –a bwtsw [path to the reference genome in FASTA format]

~bwa mem [path to indexed reference genome] [path to forward read in FASTQ format] [path to reverse read in FASTQ format] > name.sam

The final alignment created by BWA in sequence alignment map (SAM) format, was converted into binary alignment map (BAM) format with the use of SAMtools. Subsequently, the BAM files were sorted, and mapping statistics were calculated:

~samtools view –bS name.sam > name.bam

~samtools sort name.bam –o sorted_name.bam

~samtools flagstat sorted_name.bam

In order to find the most accurate reference genome, for *C. jejuni* isolates which showed a percentage of mapped reads below 90% against *C. jejuni* NCTC 11168 (NC_002163.1), additional reference-based mapping was proceeded. Following reference genomes were used: *C. jejuni* TS1218 (NZ_CP017860.1), *C. jejuni* RM1221 (NC_003912.7), *C. jejuni* 81-176 (NC_008787.1), *C. jejuni* RM1285 (NZ_CP012696.1), *C. jejuni* M1 (NC_017280.1) and *C. jejuni* WP2202 (NZ_CP01742.1). For *C. coli*, additional mapping was conducted for all isolates against *C. coli* ASM202418v1 (NZ_CP019977.1), *C. coli* FB1 (NZ_CP011015.1), *C. coli* 15-537360 (NC_022660.1) and *C. coli* RM5611 (NZ_CP007179.1).

To obtain an overview of the main properties of the alignment data, including genome coverage, a QualiMap v 2.0 was used (Okonechnikov, Conesa, & García-Alcalde, 2015).

2.6 Phylogenetic approaches of sequenced data

2.6.1 Core genome MLST (cgMLST)

Determination of cgMLST of analysed *C. jejuni* and *C. coli* was performed in Ridom SeqSphere+ v 3.2.1 (Ridom GmbH, Münster, Germany) at NIPH facility, based on the cgMLST scheme developed by (Cody et al., 2017). In order to determine the cgMLST target genes, a cgMLST Target Definer function was used with default parameters for gene by gene comparison.

For estimating the allelic differences between analysed 40 *Campylobacter* isolates, their genomes were first assembled in Velvet v 1.1.04 (Zerbino & Birney, 2008), and subsequently imported into Ridom SeqSphere+. These query genomes were run through BLASTn search against cgMLST targets. The parameters for this procedure were as follows: identity percentage: 80%, aligned: 100%, word size: 11, mismatch penalty: -1, match reward: 1, gap open costs: 5 and gap extension costs: 2. The good targets were accepted if they were present in ≥95% analysed genomes.

2.6.2 Core genome tree

De novo assembled contigs for all isolates were run through Prokka (Seemann, 2014), a software for bacterial, archaeal and viral genomes annotation. Subsequently, annotated assemblies in the GFF3 format were run through Roary (Page et al., 2015), a pipeline for calculating the pan and core genome. The command line for this program was as follows:

~roary –f –e –n –v ./[folder including files in GFF format]/*.gff

Minimum percentage identity for BLASTp (a part of Roary pipeline) was set to 95%, whereas a percentage to define a gene to be core was 99%. Option –e was proceeded to create a multiFASTA alignment of core genes using PRANK (Löytynoja, 2014), which in turn could be used in building a phylogenetic tree. Subsequently, a multiFASTA alignment was converted to nexus format, compatible with PAUP* (Phylogenetic Analysis Using Parsimony) (Swofford, 2003). Core genome alignment of 40 isolates was analysed in three different phylogeny software: PAUP* v 4.0a, CLC genomic workbench v 11.0 (Qiagen, Denmark) and MEGA X (Kumar, Stecher, Li, Knyaz, & Tamura, 2018), and subsequently NJ and ML trees were constructed with default parameters. For all trees, bootstrapping was performed for estimation of phylogenetic trees. Overview of distance methods of measure and amount of bootstrap replicates used in different phylogeny tools is visualized in Table 4.

Table 4. **Overview of methods and measurement techniques used for phylogenetic approach**.

	PAUP*		CLC genomic workbench		MEGA	
	NJ	ML	NJ	ML	NJ	ML
Substitution model	Jukes-Cantor	Jukes-Cantor	Jukes-Cantor	Jukes-Cantor	Jukes-Cantor	Jukes-Cantor
Bootstrapping replications	100	100	1000	1000	1000	500

2.6.3 SNP tree

Identification of core genome orthologous SNPs was proceeded in a Synergised Pipeline for Analysis of Next-generation sequencing Data in Linux – SPANDx v 3.2 (Sarovich & Price, 2014). SPANDx was executed for raw NGS reads, which were mapped to reference genome *C. jejuni* NCTC 11168. The command was as follows:

~SPANDx.sh –r [reference genome in FASTA format] –m yes –i yes –a yes –v GCA_000009085.1.21

For increased resolution of phylogenetic approach, a merged SNP-indel matrix was created with the use of a **MergeSnpIndel.sh** script. Subsequently, the matrix was executed in PAUP* for construction of SNP trees using NJ, ML and MP methods.

2.7 Detection of virulence and antimicrobial resistance genes based on WGS data

Contigs assembled by SPAdes (see section 2.5.2) were uploaded in batch to Bacterial Analysis Pipeline (BAP) (Thomsen et al., 2016) in order to analyse bacterial genomes with default threshold for ResFinder (90%). An additional search for antimicrobial resistance genes was done in the Comprehensive Antibiotic Resistance Database – CARD, where sequences of all 40 *Campylobacter* isolates were analysed in Resistance Gene Identifier v 4.0.3 (McArthur et al., 2013) with default parameters. The genes of significance were characterized by identity, and length covering ≥ 95% of reference sequence.

A search of virulence genes was proceeded in CLC Genomic Workbench v 11.0, based on the database for virulence genes in *Campylobacter*, obtained from Virulence Factors of Pathogenic Bacteria platform (VFDB) (Chen et al., 2005). *De novo* assembled contigs from SPAdes were used in BLASTn as queries against a database including a total of 134 core chromosome and plasmid-encoded virulence genes for *Campylobacter*. All parameters for BLAST search were default (word size: 11, match reward: 2, mismatch penalty: -3, gap open

costs: 5, gap extension costs: 2, number of threads: 56). The threshold for BLASTn were set to ≥ 80% coverage and ≥ 90% identity.

3. Results

3.1 Illumina Sequencing Data Sets

A total of 40 sets of paired-end reads generated by the Illumina sequencer were obtained in FASTQ format. An average number of reads was 704731.4 bp for both, *C. jejuni* and *C. coli*. Mean coverage was calculated for both species separately, considering two reference genomes (*C. jejuni* NCTC 11168 and *C. coli* CVM N29710) and was 146.6 for *C. jejuni* and 112.4 for *C. coli* (see details in Table 5).

Table 5. **Summary of sequencing statistics for *C. jejuni* and *C. coli* (marked with a star symbol) obtained by the Illumina sequencing platform**. Mean coverage was determined considering *C. jejuni* NCTC 11168 and *C. coli* CVM N29710 as reference genomes.

Isolate	Size of file (MB)	Number of reads (bp)	Total read count	Mean Coverage
16EP001088	1117.3	1163576	2276353	247.9639
16EP000265	1173.7	1209057	2313382	246.6801
16EP001139	1157.3	1225005	2437338	244.2873
15EP000253	1184.2	1047337	1760599	224.4256
1109-1180	1163.7	1199636	2005096	203.6925
16EP000145	902.1	915620	1763028	198.0793
14EP001617	1280.8	1120259	1598679	197.3899
1109-1179	1029.0	1000448	1662301	193.3441
13EP002420	818.5	804217	1504418	187.7140
1109-1207	832.3	681757	1266839	181.4305
14EP001642	1057.7	1120259	1389629	181.2477
12EP000408	809.2	800091	1500555	179.6785
16EP000713*	889.4	891333	1620150	175.0565
13EP002426	754.6	743027	1385594	174.9723
13EP001161	760.5	745256	1379118	174.9274
13EP002546	708.4	669402	1286747	169.0184
13EP002526	737.3	714081	1281138	162.2214
13EP000133	762.0	732421	1330452	156.3077
12EP001377*	716.9	593624	1105524	150.8287
13EP001259	678.2	554134	1037563	147.7276
13EP002423	646.8	643045	1206733	145.3955
12EP000401	760.9	748654	1408631	137.4925
1109-1130	699.1	685146	1180456	135.0537
13EP001978	693.3	683157	1248733	134.6538
1109-1129	669.8	647023	1023056	117.5531
17EP001093	531.3	538988	1034103	112.4842
17EP001087	549.4	533536	899761	103.9243
16EP002233	496.0	482626	899649	99.9267
15EP000596*	537.4	472957	749663	97.5503
15EP000113	515.5	413660	666837	96.9998
15EP002192	471.7	448490	788220	95.8729
14EP000043*	505.3	492423	821312	89.8422
15EP001566	468.3	452858	780475	89.1887
17EP001113	431.6	428513	796281	88.6369
17EP001096	412.7	400962	675224	78.1389
16EP001848	384.7	370175	691523	68.2395
14EP001612	475.0	418244	521593	61.8679
13EP000100	727.3	720038	1327211	55.5449
16EP001980*	241.1	235545	441734	48.8038
14EP000843	454.8	442674	735921	40.1217
MEAN	730.1275	704731.4	1245040.5	142.3571

Trimming of the reads improved their quality, as it removed adapters and/or low-quality data (see Figure 11 and supplementary files S2–S3). Those enhanced reads were subsequently used for *de novo* assembly and reference-based mapping.

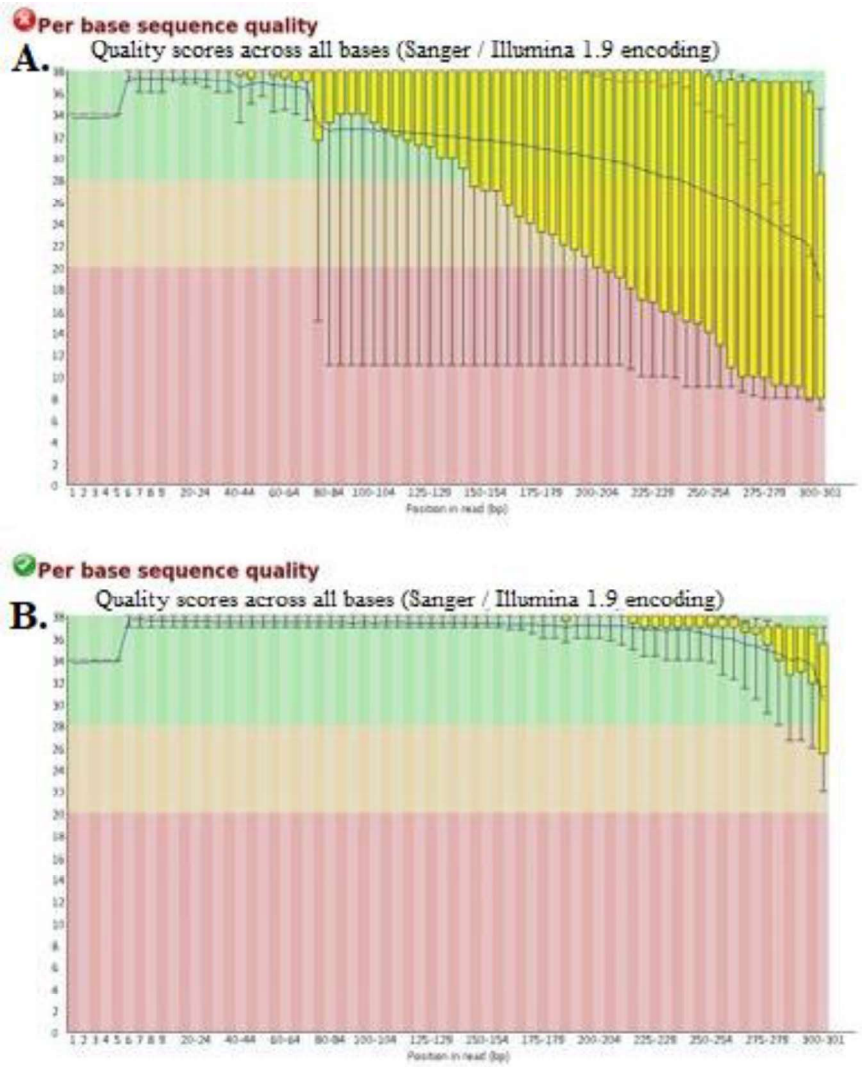

Figure 11. **FastQC visualization of per base sequence quality of isolate 14EP001617 before (A) and after trimming (B) of adapters remains and/or low quality reads.**

3.2 Assessment of whole genome assemblies

3.2.1 *De novo* assembly

Results of *de novo* assembly by SPAdes and A5-pipeline were visualized with QUAST tool, which gives insights into many matrices for analysis of genome assemblies. For calculation of *de novo* assembly statistics, *C. jejuni* NCTC 11168 and *C. coli* CVM N29710 reference genomes were chosen. Overall, SPAdes gave better assembly statistics, compared to A5-pipeline. Supplementary file S4 presents the detailed statistics of genome assembly by

SPAdes and A5-pipeline, while Table 6 shows a general data concerning genome assembly by SPAdes.

Table 6. **Mean ± Standard deviation of SPAdes' *de novo* assembly of 35 *C. jejuni* and 5 *C. coli* isolates, visualized by QUAST**. All statistics were based on contigs length ≥1000 bp.

	C. jejuni	*C. coli*
Number of contigs	19.64±10.35	22.14±11.40
Largest contig size	513±146.9 kbp	412.86±128 kbp
A total length of the assembly	1.63±0.06 Mbp	1.71±0.04 Mbp
Length of the reference	1.64 Mbp	1.73 Mbp
N50	215757±56316 bp	210445±49387 bp
GC%	30.46±0.22	31.35±0.13
Misassemblies		
Number of misassemblies	14.64±10.26	14.28±5.8
Number of misassembled contigs	6.05±2.95	6.28±1.66
Misassembled contigs length	1.221±0.39 Mbp	1.15±0.32 Mbp
Mismatches		
Number of mismatches per 100 kb	1151±814	755±240
Number of indels per 100 kb	37.61±35.52	31.25±8.8
Number of N's per 100 kb	0	0
Genome statistics		
Genome fraction (%)	88.28±20.84	91±3.1
Duplication ratio	1.06±0.03	1.05±0.02
NGA50	120825±55435	110547±58875

Use of PlasmidSPAdes enabled to detect plasmids for all strains in this study, and subsequent BLASTn searches found sequence homologs of those plasmids in respective *de novo* assembled contigs. Further investigation of plasmids in BLASTn against publicly available databases showed their inhomogeneous relativeness (see Table 7). Less than half of queried plasmids that passed the search filters, showed close homology to actual plasmids (shaded in grey) while remaining plasmids obtained the best alignments with chromosomes. Some predicted plasmids (very small size and most likely false positive) presented high homology with a p301-4 plasmid of *E. coli* DH5alpha, whereas others were homologous to chromosomes of species like *Sphingorhabdus* or *Staphylococcus* (see supplementary file S5). Interestingly, a plasmid found in the genome of *C. jejuni* isolate 16EP001139 was homologous with *C. coli* plasmid; and plasmid predicted for *C. coli* isolate 15EP000596 presented homology with a pTet plasmid of *C. jejuni* strain S3.

Table 7. **BLASTn searches for plasmids as queries against publicly available NCBI database.** Homology to plasmids belonging to *Campylobacter* species is shaded in grey. Isolates marked with a star (*) are *C. coli.*

Isolate	Query length [bp]	Best BLAST hit	Hit length [bp]	Query cover [%]	E-value	Identity [%]	Accession number
16EP002233	21887	*C. jejuni* strain 104 chromosome	1558306	97	0.0	97	CP023343.1
1109-1130	26861	*C. jejuni* RM1246 ERRC plasmid	45197	95	0.0	98	CP022471.1
1109-1179							
1109-1180							
1109-1207							
13EP000133	34318	*C. jejuni* strain 00-6200 chromosome	1670781	87	0.0	99	CP010307.1
16EP001848	34337	*C. jejuni* strain 00-0949 chromosome	1745537	98	0.0	98	CP010301.1
12EP001377*	35352	*C. coli* plasmid pCC31	44707	95	0.0	99	AY394560.1
14EP000043*	35651	*C. coli* strain CVM N29710 chromosome	1673221	90	0.0	99	CP004066.1
14EP000843*	36826	*C. coli* strain ZV1224 chromosome	1837306	96	0.0	98	CP017875.1
13EP000100*	38347	*C. coli* strain FB1 chromosome	1658607	88	0.0	99	CP011015.1
15EP000596*	40606	*C. jejuni* strain S3 plasmid pTet	43222	99	0.0	98	CP001961.1
16EP001139	46347	*C. coli* strain FB1 plasmid pFB1TET	44826	82	0.0	95	CP011017.1
13EP002420	48995	*C. jejuni* strain 12567 chromosome	1705686	100	0.0	100	CP028909.1
13EP002546							
1109-1129	49132	*C. jejuni* NCTC 12660 chromosome	1684042	99	0.0	99	CP028910.1
13EP001259	50483	*C. jejuni* strain 81-176 plasmid pTet	45210	85	0.0	99	AY714214.1
17EP001087	57376	*C. jejuni* strain OD267 plasmid pCJDM67 L	116883	98	0.0	99	CP014745.1
14EP001642	61254	*C. jejuni* strain 81116 chromosome	1628115	99	0.0	99	CP000814.1
17EP001096	68927			100			
12EP000401	274123	*C. jejuni* strain CJ677CC523 chromosome	1667224	100	0.0	99	CP010508.1

3.2.2 Reference-based mapping

Because of the genomic diversity of *Campylobacter* species, finding a reference genome which was utmostly related to analyzed *C. jejuni* and *C. coli*, respectively, was crucial for obtaining adequate reference-based mapping statistics for both species. In the analysed data sets, the best overall numbers were obtained for *C. jejuni* NCTC 11168 and *C. coli* CVM N29710.

Pair-ended sequencing reads of 35 *C. jejuni* and 5 *C. coli* isolates were mapped in BWA to their reference genomes *C. jejuni* NCTC 11168 and *C. coli* CVM N29710, respectively. The assessment of reference-based mapping was made based on the statistics presented as a percentage of mapped reads and percentage of mapped reads in pair. For most of *C. jejuni* isolates, the percentage of mapped reads to the chosen reference genome reached more than 90% and for the remaining isolates, it did not drop lower than 78.4% (Table 8). Isolates 13EP000100 and 14EP000843, on the other hand, appeared to be definite outliers in the analysed group, as only 49.4% and 59.8% of their reads, respectively was mapped to *C. jejuni* NCTC 11168 genome. Because of low statistics for these isolates, their reads were mapped to 6 different *C. coli* reference genomes (described in section 2.5.3), to select the appropriate reference genome. This improved mapping considerably, as a total of 75.54% reads for isolate 13EP000100, and 83.85% reads for isolate 14EP000843, were mapped to *C. coli* CVM N29710. These two isolates have shown more similarities with *C. coli* than *C. jejuni*, what explained their poor statistics of reference-based mapping. In further analysis, these isolates were considered as *C. coli*.

Within remaining *C. coli* isolates, the reads of all isolates were mapped to the reference genome in ≥93%. For details, see supplementary file S6 which presents reference-based mapping results for all analysed *Campylobacter* isolates.

Table 8. **Statistics for reference-based mapping by BWA**. *C. jejuni* and *C. coli* isolates were mapped to *C. jejuni* NCTC 11168 and *C. coli* CVM N29710 reference genomes, respectively. The values in the brackets for isolates 13EP000100 and 14EP000843 (marked with a star) refer to mapping statistics against *C. coli* CVM N29710.

Pathogen	Isolate ID	% of Mapped Reads	% of Reads in Pair
C. coli	15EP000596	99.11	98.70
C. jejuni	13EP002546	98.70	97.97
C. jejuni	13EP001161	98.69	98.24
C. jejuni	13EP002426	98.56	98.17
C. jejuni	13EP002420	98.40	97.75
C. jejuni	13EP002526	98.18	97.56
C. jejuni	13EP002423	98.15	97.21
C. jejuni	16EP000145	95.88	94.52
C. coli	12EP001377	95.77	95.31
C. jejuni	12EP000408	95.38	94.86
C. coli	16EP001980	94.20	92.77
C. jejuni	15EP000113	94.05	94.82
C. coli	16EP000713	93.79	92.69
C. jejuni	16EP001088	93.76	92.44
C. coli	14EP000043	93.71	92.67
C. jejuni	14EP001612	93.40	92.92
C. jejuni	15EP000253	93.31	92.64
C. jejuni	14EP001617	93.21	92.53
C. jejuni	17EP001093	93.14	91.74
C. jejuni	13EP001259	92.50	92.22
C. jejuni	1109-1207	92.46	92.02
C. jejuni	14EP001642	92.26	91.48
C. jejuni	17EP001096	92.21	91.28
C. jejuni	1109-1179	92.09	91.30
C. jejuni	16EP000265	92.05	90.55
C. jejuni	17EP001113	91.94	90.84
C. jejuni	15EP002192	91.85	91.41
C. jejuni	13EP000133	91.52	90.55
C. jejuni	16EP001139	90.91	89.13
C. jejuni	17EP001087	90.53	90.03
C. jejuni	15EP001566	90.38	89.49
C. jejuni	1109-1130	90.37	89.92
C. jejuni	1109-1129	90.15	89.68
C. jejuni	13EP001978	87.73	86.61
C. jejuni	16EP002233	87.41	86.63
C. jejuni	1109-1180	84.44	86.98
C. jejuni	12EP000401	81.75	80.36
C. jejuni	16EP001848	78.38	77.23
C. jejuni	14EP000843*	59.76 (83.85)	57.66 (82.52)
C. jejuni	13EP000100*	49.43 (75.54)	47.30 (74.00)

3.3 Phylogenetic analysis

Contigs generated in *de novo* assemblers were used for creating a cgMLST scheme and for distinguishing the core genome of *Campylobacter*. Subsequently, phylogenetic relationships between isolates were described, based on core genome alignment. In parallel, phylogeny based on SNP was performed, using raw reads from Illumina sequencing. The evolutionary trees were constructed using UPGMA, NJ, ML and MP approaches, and are presented in sections 3.3.1 – 3.3.3.

Independently of phylogeny algorithm used, all trees resulted in presenting distance between *C. jejuni* and *C. coli* isolates. This distance, described as a number of substitutions per site, can be observed in Figure 12, which shows a phylogram based on core genome of the 40 *Campylobacter* isolates, created in PAUP* with NJ method.

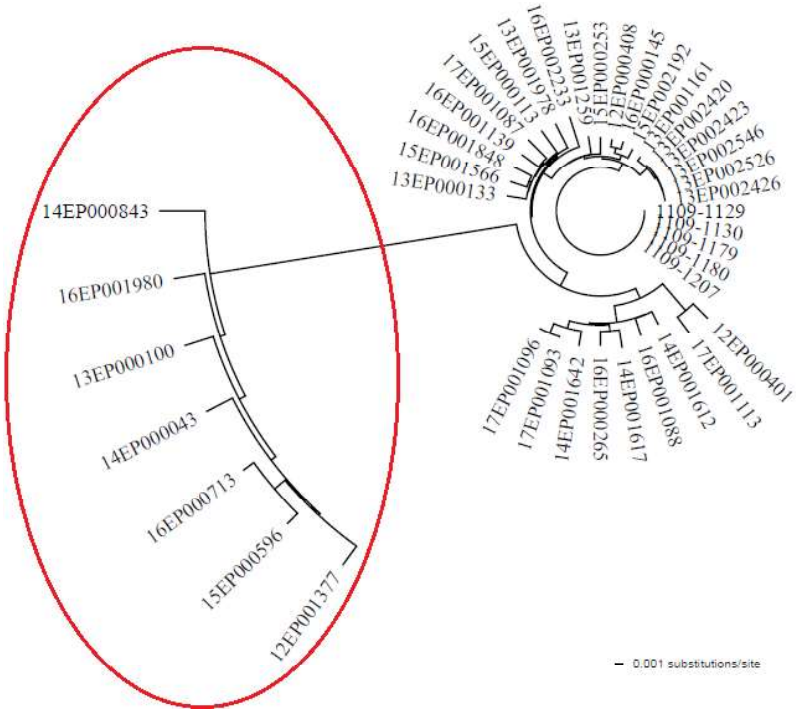

Figure 12. **Phylogram representing the distance between *C. coli* (circled in red) and *C. jejuni*.**

3.3.1 cgMLST

With the use of this cgMLST scheme, the ST types were recognized for all of the analysed isolates. A detailed overview of target genes list and cgMLST results are included in supplementary file S7. Relationships between the isolates were visualized with UPGMA (Figure 13) and minimum spanning (Figure 14) trees, based on the pairwise comparison of allelic profiles. *Campylobacter* isolates were of different ST types; however, three clusters

characterized by ≤ 2 allelic differences could be identified (Table 9). All isolates described as clusters were identified as *C. jejuni,* and all of them were acquired in Norway. Isolates belonging to each cluster were recovered in the same year – 2011, 2013 and 2017 for clusters A, B, and C, respectively.

Table 9. **Description of clusters determined from the cgMLST approach, based on ≤ 2 allelic differences.**

Isolate	Year of collection	Cluster	Allelic difference	ST type
1109-1129, 1109-1130, 1109-1179, 1109-1180, 1109-1207	2011	A	0	21
13EP002420, 13EP002423, 13EP002426, 13EP002526, 13EP002546	2013	B	1	53
17EP001093, 17EP001096	2017	C	2	583

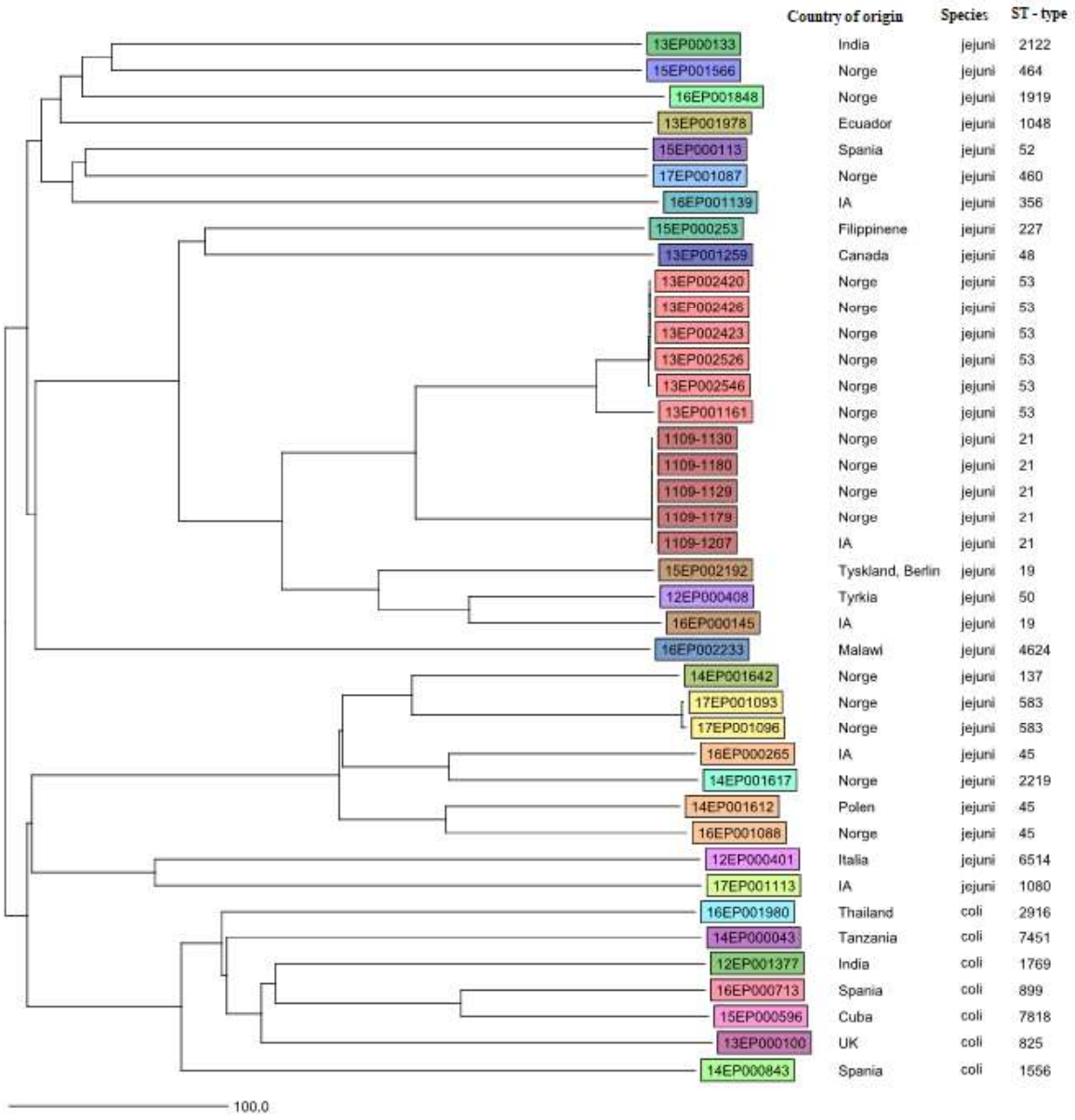

Figure 13. **UPGMA tree based on 637 target core genes established by the cgMLST scheme**. Total of 33 *C. jejuni* and 7 *C. coli* isolates were described with regards to country of origin and ST type and are coloured according to their ST types.

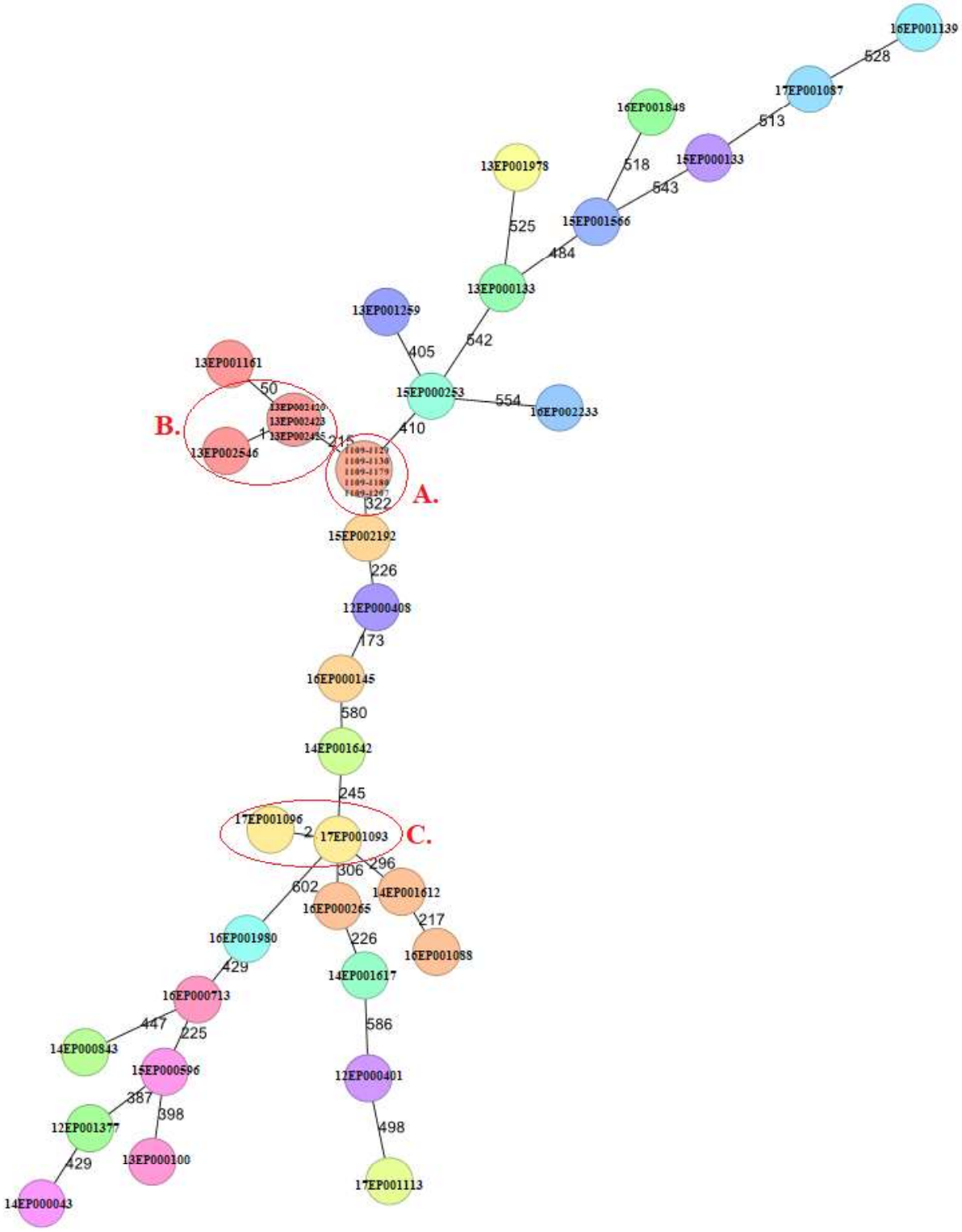

Figure 14. **Minimum spanning network tree built from the core genome of 40 *Campylobacter* isolates**. The isolates are coloured according to the ST type. Numbers between isolates describe the amount of allelic differences between them. The lengths of lines are not proportional to the numbers. Three clusters (circled in red) are distinguishable, with allelic difference ≤ 2.

3.3.2 Phylogeny based on core genome

The core genome alignment generated with Roary was analysed in three different software to find the best quality of phylogenetic relationships between taxa. Phylogeny was built for *C. jejuni* and *C. coli*. Overall, trees generated in PAUP* and MEGA presented similar topology, whereas CLC Genomic Workbench gave slightly different results. Phylogenetic trees obtained by PAUP* are presented here to visualise the evolutionary distances between *Campylobacter* isolates. Figures acquired in MEGA and CLC are included in supplementary file S8.

The NJ tree created in PAUP* was based on core genome alignment of 40 taxa, consisting of 234429 nucleotides. The evolutionary distances (i.e., where two sequences differed) between analysed *Campylobacter* isolates were calculated in CLC and are marked in red in Figure 15. Based on core genome analysis, three clusters could be defined. Isolates 1109-1129, 1109-11330, 1109-1179, 1109-1180 and 1109-1207 were clustered together as there were no nucleotide differences between them. Differences between isolates 13EP002423, 13EP002546, 13EP002420, 13EP002426 and 13EP002526 were ≤ 2 and there were 22 nucleotide differences between 17EP001096 and 17EP001093.

None of the seven *C. coli* isolates were grouped into one cluster. There were 3806±1487 nucleotide differences between analysed *C. coli* isolates, and 21434±452 differences between *C. coli* and corresponding neighbour *C. jejuni* isolates. Within *C. jejuni*, no specific outlier was determined. Interestingly, sister taxa were not always grouped by the year of collection. Many isolates were related to isolates retrieved in up to 5 years' interval (for example isolate 12EP000401 and 17EP001113).

Similar results were obtained by ML approach (Figure 16). The dissimilarities between NJ and ML trees were laying on branching of six *C. coli* isolates 14EP000843, 13EP000100, 16EP000713, 15EP000596, 14EP000043 and 12EP001377. Moreover, the ML tree presents isolates 13EP001978 and 13EP000133 as sister taxa, while NJ tree assumes that they are more distant. Differences were also noticed considering branching for isolate 16EP002233.

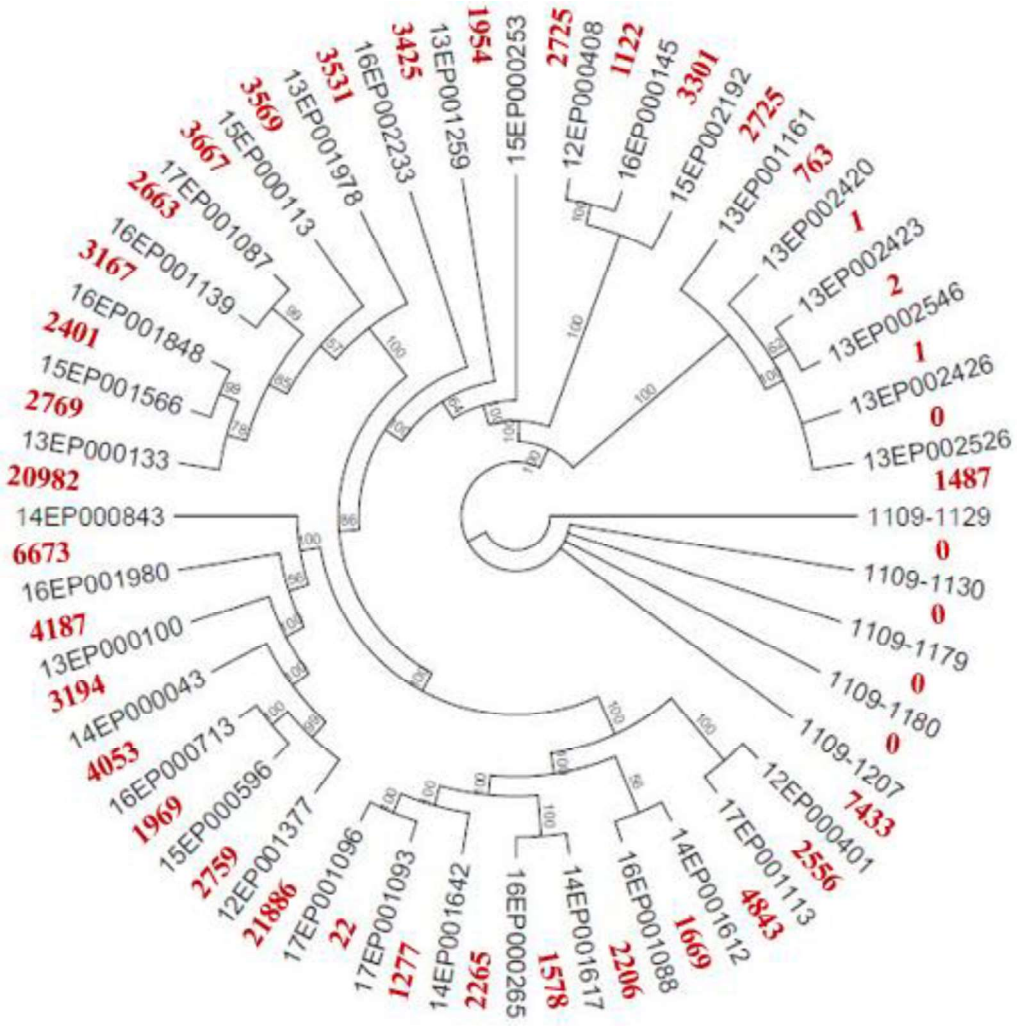

Figure 15. **Unrooted bootstrap consensus tree constructed with neighbour joining method, based on core genome assembly of 40 *Campylobacter* isolates.** The NJ tree was estimated in PAUP* with distance as minimum evolution and Jukes-Cantor model for DNA distances. All positions containing gaps and missing data were excluded. Bootstrapping was proceeded with the NJ method, 100 replicates, and values lower than 50% are not shown in the figure. Pairwise nucleotide comparison between neighbouring isolates was calculated in CLC Genomic Workbench and are given in red.

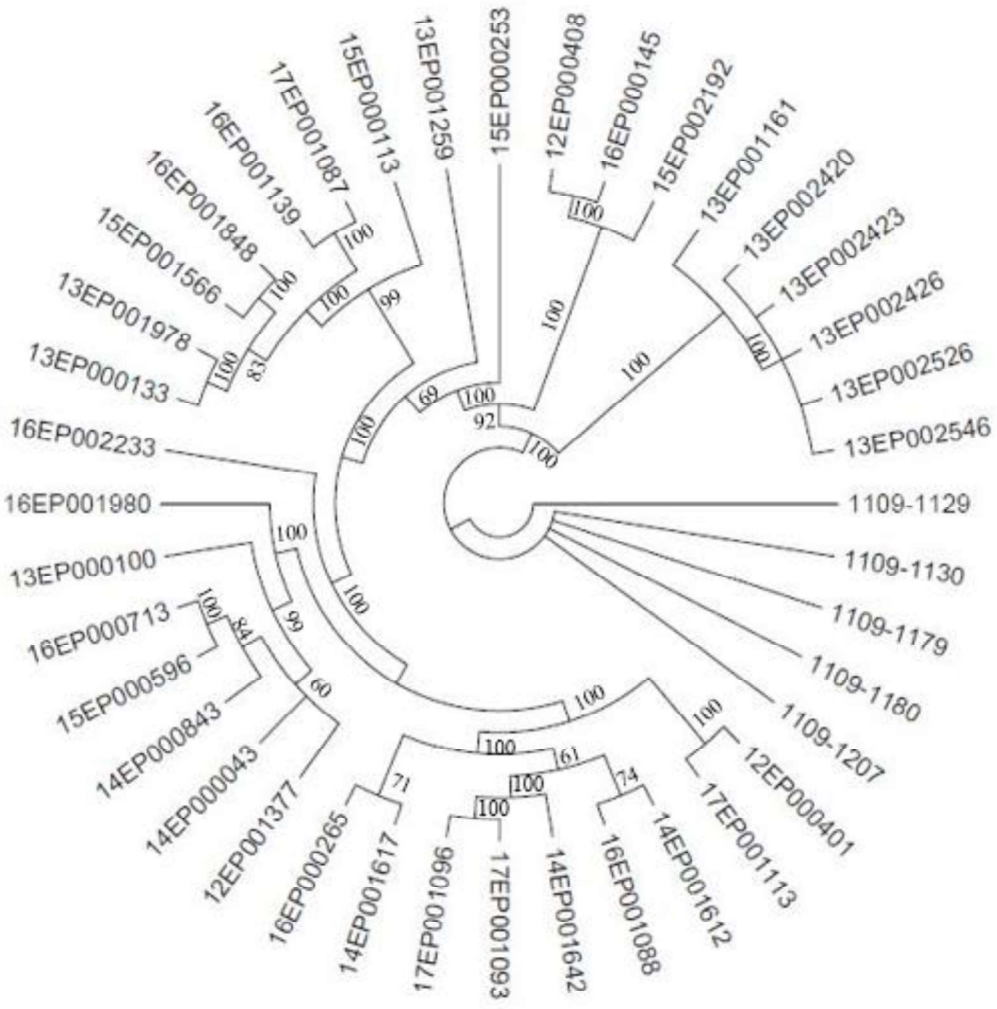

Figure 16. **Unrooted bootstrap consensus tree constructed with maximum likelihood method, based on core genome assembly of 40 *Campylobacter* isolates.** The tree was estimated in PAUP* with default parameters. All positions containing gaps and missing data were excluded. Bootstrapping was proceeded with 100 replicates, and values lower than 50% are not shown in the figure.

3.3.3 Phylogeny based on SNP

Phylogenetic analysis based on SNPs from whole genomes of sequenced *C. jejuni* and *C. coli* was performed in PAUP*, using an indel_SNP_matrix.nex file obtained from SPANDx. The details on a number of SNPs passing filters established in SPANDx are included in supplementary file S9. NJ, ML and MP trees based on 65506 SNPs/indels presented comparable topology, with variations in branching for *C. coli* species. The pairwise SNP differences between the isolates were calculated in PAUP* and are marked in green in Figure 17. Analysis based on SNPs gave similar results to those, established with cgMLST

and core genome data. Three complexes (A-C) with differences ≤ 2 were determined based on SNP data. Isolates 1109-1129, 1109-1130, 1109-1179, 1109-1180 and 1109-1207 were clustered together into complex A, based on ≤ 1 SNP difference. Complex B consists of isolates 13EP002546, 13EP002526, 13EP002426, 13EP002423 and 13EP002420, and varies by ≤ 1 SNP difference between them. Finally, cluster C includes isolates 17EP001096 and 17EP001093 with 2 SNP differences between each other.

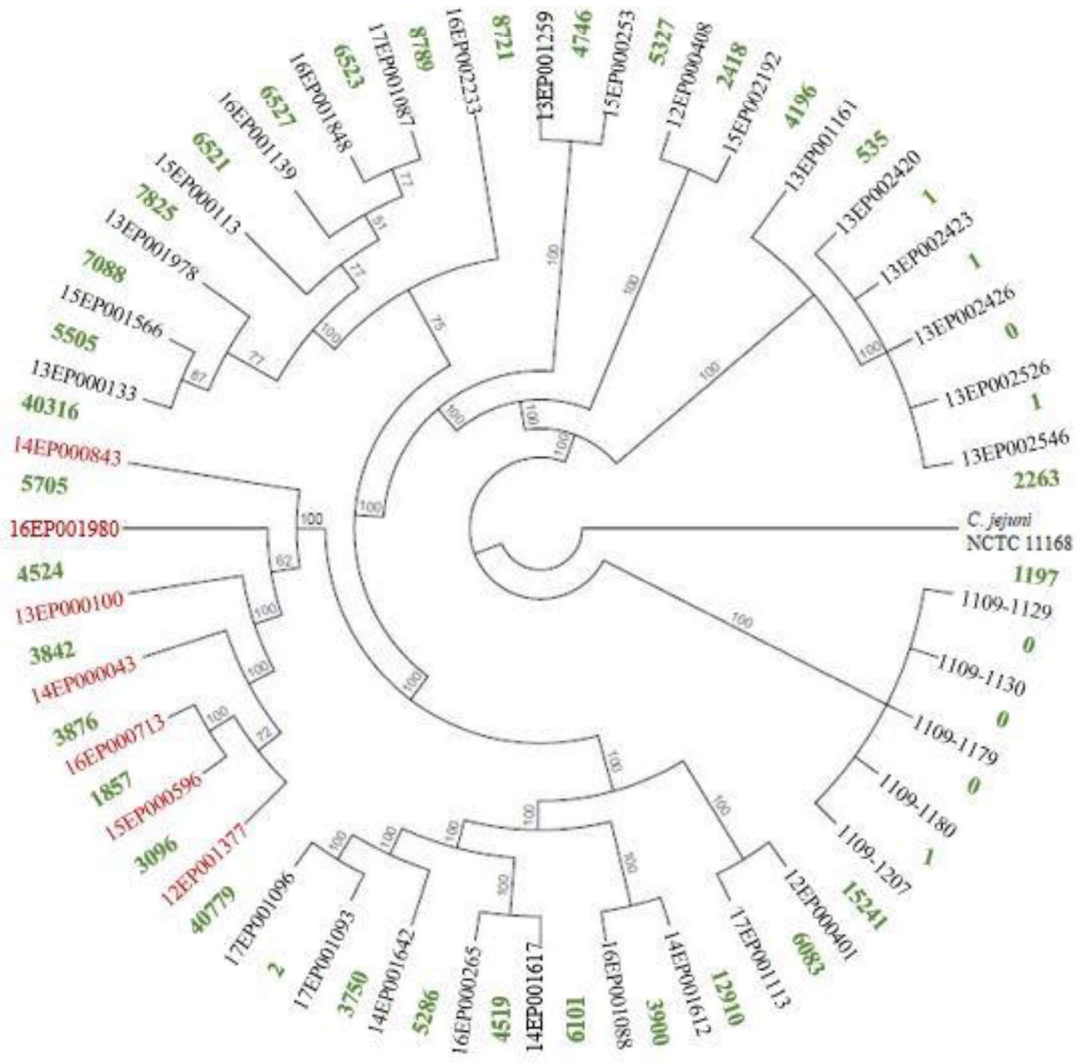

Figure 17. **Unrooted neighbour joining tree generated in PAUP* based on SNP/Indel matrices for 40 *Campylobacter* isolates and one reference genome (*C. jejuni* NCTC 11168).** Pairwise SNP differences between neighbouring isolates are marked in green. Seven *C. coli* isolates are marked in red. The NJ tree was constructed with 65506 characters (SNPs or/and indels) and with distance as minimum evolution. Bootstrapping was performed with 100 replicates using the NJ method.

C. coli isolates (marked in red) were not clustered into a complex, as there were 3817±1185 SNP differences between them. Moreover, alterations between *C. coli* and neighbouring *C. jejuni* isolates reached 40548±232 SNP differences.

ML tree shows resemblances with NJ tree, however, some variations in branching can be noticed (Figure 18). Dissimilarities in branching are noticeable for *C. coli* isolates (marked in red) as well as for isolates 13EP001978 and 16EP001848.

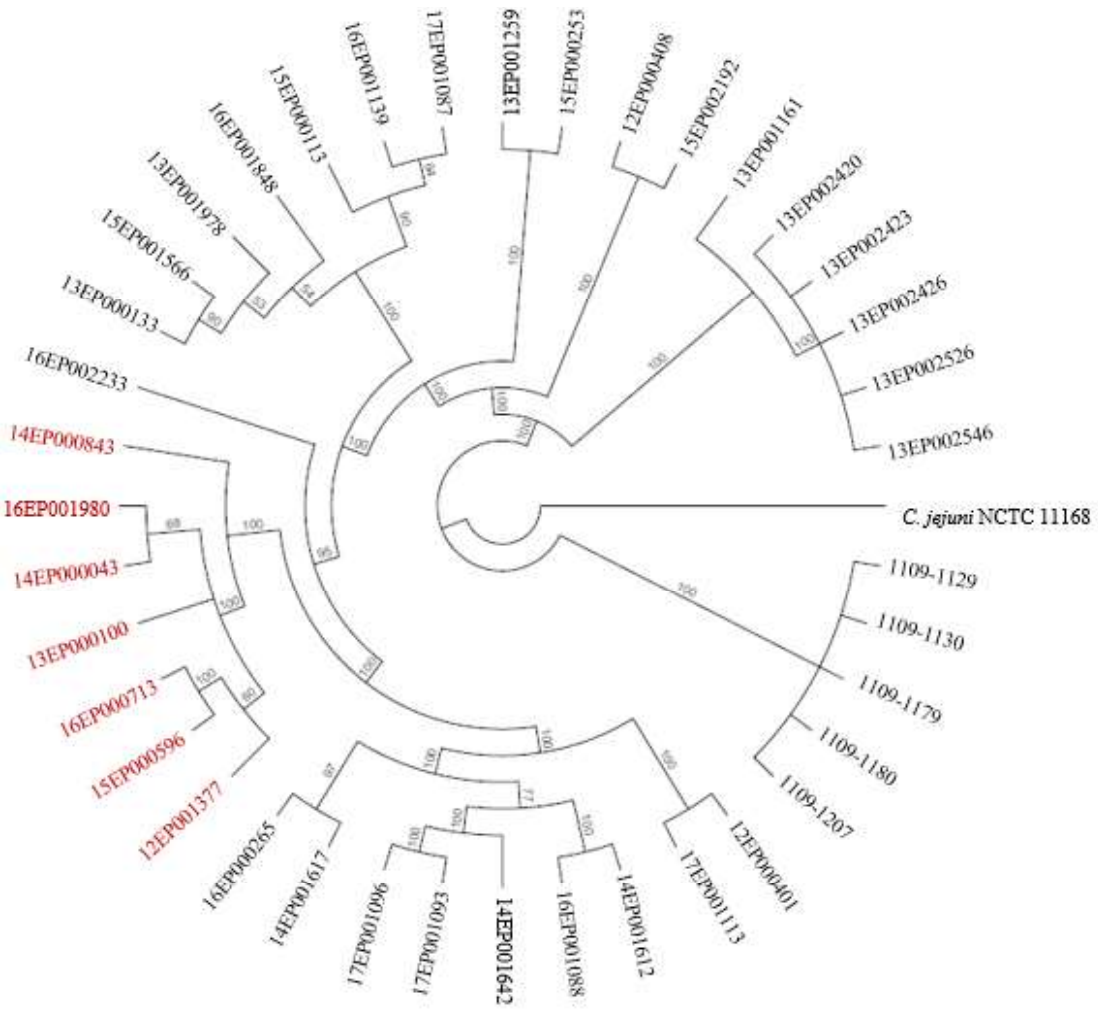

Figure 18. **Unrooted maximum likelihood tree constructed PAUP* based on SNP/Indel matrices for 40 *Campylobacter* isolates and one reference genome (*C. jejuni* NCTC 11168).** Seven *C. coli* isolates are marked in red. The tree was constructed with 65506 characters (SNPs or/and indels) and bootstrapping was performed with 100 replicates.

MP tree (Figure 19) presented less information about relationships between *C. coli* isolates and some of *C. jejuni* isolates. For example, MP tree shows polytomy for isolates 15EP000113, 16EP001139 and 17EP001087, whereas NJ and ML trees distinguished the way how those isolates are related to each other and remaining *Campylobacter*.

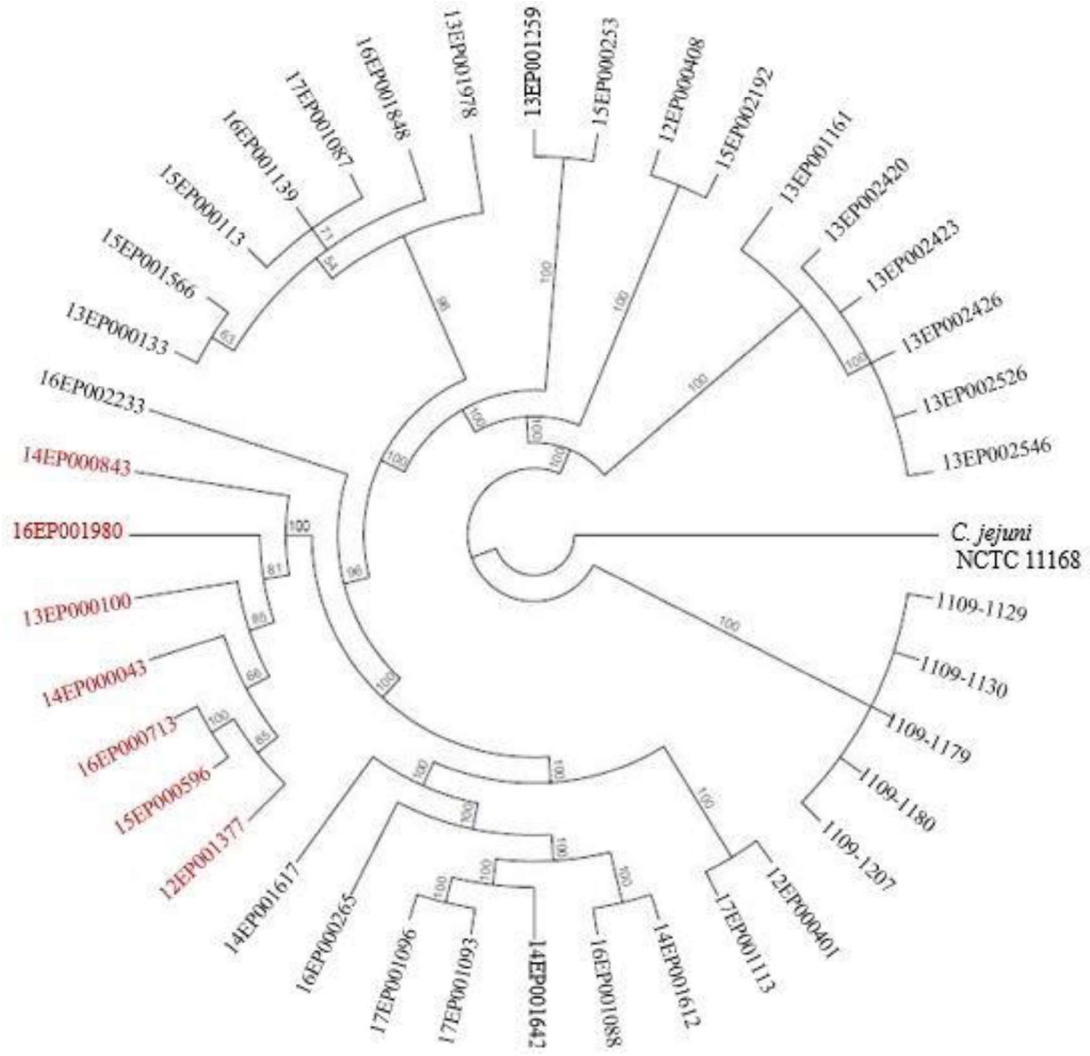

Figure 19. **Unrooted maximum parsimony tree constructed PAUP* based on SNP/Indel matrices for 40 *Campylobacter* isolates and one reference genome (*C. jejuni* NCTC 11168).** Seven *C. coli* isolates are marked in red. The tree was constructed with 65506 characters (SNPs or/and indels) and bootstrapping was performed with 100 replicates.

3.4 WGS based detection of virulence and antimicrobial resistance genes

3.4.1 Virulence genes detection

BLASTn searches against VFDB resulted in a determination of chromosomal virulence genes. The results from this study have revealed that analysed isolates carried genes enabling among others motility, chemotaxis, adherence, and invasion (data not shown). However, no plasmidal virulence genes belonging to the type IV secretion system were detected.

3.4.2 Antimicrobial resistance genes detection

A module of BAP, ResFinder, detected a total of seven resistance genes, and they were as follows: *bla*$_{oxa-61}$, *tetO, aph(2")-Ih, aph(2")-If, aph(2")-Ic, aph(3')-III,* and *aadE*. A *tetO* and *bla*$_{oxa-61}$ genes encode resistance to TET and β-lactams, respectively, while *aph(2")-Ih, aph(2")-If, aph(2")-Ic* genes are known to encode resistance to GEN. The remaining two genes encode aminoglycoside-modifying enzymes (Griggs et al., 2009; Sougakoff, Papadopoulou, Nordmann, & Courvalin, 1987; Zhao et al., 2015).

Out of 40 analysed isolates, 67.5% (n=27) appeared to have the *bla*$_{oxa-61}$ gene. Presence of genes encoding resistance against tetracycline and gentamycin was less frequent: 22.5% (9/40) and 10% (4/40), respectively. Overall, most of the isolates harboured only one gene encoding resistance to only one antimicrobial drug. However, isolates with two or more resistance genes were noted. The highest number of genes was observed in isolates 15EP000253, and 16EP001980, which carried *bla*$_{oxa-61}$, *tetO,* and *aph(3')-III*, as well as other genes encoding aminoglycoside-modifying enzymes (see supplementary file S10).

RGI searches against CARD database provided more in-depth detection of antimicrobial resistance genes. Besides genes distinguished by BAP, it also detected *cmeA, cmeB, cmeC, cmeR, tet(W/N/W), ant9-Ia, sat4, aph(2")-IIIa, aad6,* and *bla*$_{oxa-184}$ genes; however RGI did not detect *aph(2")-Ih* and *aph(2")-Ic* genes. The *cmeABC* genes encode an efflux pump which is responsible for multidrug resistance, including resistance to fluoroquinolones and macrolides (Lin et al., 2002); *ant9-Ia, sat4, aph(2")-IIIa, aad6* are genes encoding aminoglycoside-modifying enzymes; *tet(W/N/W)* and *bla*$_{oxa-184}$ are responsible for resistance against TET and β-lactams, respectively.

More than 73% of all isolates (n=29) carried operon consisting of *cmeA, cmeB,* and *cmeC* genes, along with *cmeR* – a transcriptional repressor for the efflux pump (Lin, Akiba, Sahin, & Zhang, 2005). However, isolates with resistance genotype including only one

(*cmeB*), or three (*cmeA-cmeC-cmeR*) genes from *cme* operon were observed as well (Table 10). Moreover, RGI distinguished that 75% of isolates (n=30) had genes homologous to either *oxa-61* or *oxa-184*, and 27.5% (n=11) possessed either *tetO* or *tet(W/N/W)* gene variants. Although four isolates carried genes encoding aminoglycoside-modifying genes, only three of them had genes responsible for resistance towards GEN. This result is incomplete compared to AMR genes detection by BAP, which described four isolates to carry GEN-resistance gene.

Overall, most of *C. jejuni* isolates carried genes encoding cmeABC multidrug efflux pump, the presence of which was not that common within *C. coli* isolates. What is more, the frequency of *C. coli* carrying genes encoding resistance to more than one group of antibiotics seemed to be higher, compared to *C. jejuni*.

Table 10. **AMR genotype profiles established by RGI against CARD database**.

Resistance genotype	Number of *C. jejuni* isolates	Number of *C. coli* isolates
bla_{oxa-61}	-	1
tetO	-	2
cmeA, cmeC	1	-
cmeB, tetO, bla_{oxa-61}	-	1
cmeR, cmeA, cmeB, cmeC	5	-
cmeR, cmeA, cmeC, bla_{oxa-61}	1	-
cmeB, tetO, aph(2")-If, aph(3')-IIIa	-	1
cmeR, cmeA, cmeB, cmeC, bla_{oxa-61}	18	-
cmeR, cmeA, cmeB, cmeC, $bla_{oxa-184}$	2	-
cmeR, cmeA, cmeB, cmeC, tetO	1	-
cmeR, cmeA, cmeC, $bla_{oxa-184}$, tetO	1	-
cmeR, cmeA, cmeC, bla_{oxa-61}, tet(W/N/W)	1	-
cmeR, cmeA, cmeB, cmeC, bla_{oxa-61}, tetO	2	-
aph(2")-IIIa, aph(3')-IIIa, ant(9)-Ia, sat-4, tet(W/N/W), tetO, bla_{oxa-61}	-	1
cmeB, cmeC, bla_{oxa-61}, tetO, aph(2")-If, aph(3')-IIIa, sat-4, aad(6)	-	1
cmeR, cmeA, cmeB, cmeC, bla_{oxa-61}, tetO, tet(W/N/W), aph(2")-If, aph(3')-IIIa, sat-4,	1	-

3.5 Antimicrobial susceptibility testing

The results of antimicrobial susceptibility testing performed on 33 *C. jejuni* and 7 *C. coli* isolates are presented in Table 11. Considering epidemiological cut-offs, 35% (14/40) of *Campylobacter* were described as susceptible wild type, whereas remaining 65% (26/40) as non-wild type with reduced susceptibility. Within *C. jejuni* predominance of resistance was detected to ERM (27%), followed by resistance to CIP (24%), NX (24%) and TET (21%), to a less extent to GEN (12%). As for *C. coli*, most of the isolates were resistant to CIP (85%), TET (85%), NX (85%), followed by resistance to ERM (71%) and GEN (42%). The MICs for CIP-, ERM- and TET-resistant isolates were greatly higher than clinical breakpoints distinguished for these antimicrobial drugs by EUCAST, reaching the values of 32 μg/ml for CIP, and 256 μg/ml for TET and ERM.

Overall, 26/40 isolates were resistant to one or more antimicrobial drug, 11 isolates were resistant to only one antimicrobial agent, and 5 isolates were resistant to two tested antibiotics. Multiresistance was described as resistance to three or more antimicrobial drugs and was found within 9 isolates, the 4 of which were resistant to all tested antibiotics. Overall, considering percentage, *C. coli* presented higher resistance tendency to all tested antimicrobial drugs, and a higher tendency to multiresistance compared to *C. jejuni*. The multiresistant isolates were recovered between 2012 and 2016 with no specific time points. Regarding country of origin, 7 multiresistant isolates were imported to Norway from India, Ecuador, Spain, Philippines or Thailand, while two were acquired domestically.

Table 11. **Overview on antimicrobial susceptibility testing results**. Table presents values of MICs for tested antimicrobial drugs. Values highlighted in green refer to isolates which are susceptible, whereas those, marked in orange refer to isolates resistant to a specific antibiotic. *C. coli* isolates are marked by stars.

Isolate	CIP	TET	ERM	GEN	NX	MR
13EP000100*	32	256	256	256	256	R
16EP001980*	32	256	256	256	256	R
16EP000713*	32	256	256	256	256	R
15EP000253	32	128	256	256	256	R
14EP000843*	32	256	256	2	256	R
13EP001978	32	256	8	1	256	R
15EP001566	32	256	8	0.5	256	R
12EP001377*	32	128	256	2	256	R
16EP001848	32	16	256	0.25	256	R
16EP000145	32	0.125	2	0.5	256	S
13EP000133	32	0.5	2	0.5	256	S
15EP000113	32	0.125	2	0.5	256	S
14EP000043*	32	0.25	0.25	1	256	S
12EP000408	8	0.125	1	0.5	256	S
15EP000596*	0.25	256	1	0.5	8	S
13EP001259	0.125	16	0.5	0.5	4	S
14EP001612	0.25	8	0.5	2	0.25	S
12EP000401	0.125	4	2	1	16	S
16EP002233	0.25	0.25	256	0.5	4	S
16EP000265	0.5	0.125	32	0.5	4	S
15EP002192	0.25	0.25	8	1	4	S
16EP001088	0.125	0.25	8	1	4	S
16EP001139	0.25	0.25	8	0.5	8	S
14EP001617	0.5	0.5	2	8	0.5	S
14EP001642	0.25	0.25	2	4	1	S
13EP001161	0.25	0.25	4	4	16	S
1109-1129	0.125	0.125	0.5	0.5	4	S
1109-1130	0.125	0.125	0.5	0.5	2	S
1109-1179	0.125	0.125	0.5	0.5	4	S
1109-1180	0.125	0.125	1	0.5	4	S
1109-1207	0.125	0.125	1	0.5	2	S
13EP002420	0.125	0.125	1	0.25	4	S
13EP002423	0.125	0.125	1	0.25	8	S
13EP002426	0.125	0.125	1	0.5	4	S
13EP002526	0.125	0.125	1	0.5	4	S
13EP002546	0.125	0.125	1	0.5	4	S
17EP001087	0.125	0.25	0.5	0.5	4	S
17EP001093	0.125	0.5	1	0.5	8	S
17EP001096	0.125	0.25	2	0.5	4	S
17EP001113	0.5	0.25	2	0.5	8	S

CIP- ciprofloxacin; TET- tetracycline; ERM- erythromycin; GEN- gentamicin; NX- nalidixic acid; MR- multiresistant isolates (resistant to three or more antibiotics); S- susceptible; R- resistant.

4. Discussion

4.1 Downstream analysis of sequence data

Quality and coverage of sequenced data are crucial in WGS data analysis. Short read lengths and errors produced by NGS platforms might impede a correct genome assembly. Therefore, an increase in the number of reads followed by an increase in coverage depth is required to overcome this problem (Sims, Sudbery, Ilott, Heger, & Ponting, 2014).

The FastQC reports indicated that the mean quality of reads after trimming was good for all sequenced isolates. However, per base sequence content reached the differences more extensive than 20% within the beginning and the end of the reads (see details in supplementary file S3). The possible solution to this issue might be a manual removal of bases, where the conflict was noticed. The sequencing depth towards the reference genomes was abundant, ranging from 40 to 247X (Table 5). Therefore, we believe that most of the sequenced genomes were covered sufficiently to ensure the assembly quality.

Subsequent evaluation of the genome assembly in QUAST showed that the total lengths of the assemblies were similar to lengths of the reference genomes (Table 6 and supplementary file S4). Moreover, a relatively low number of contigs was obtained (average 20), what might suggest that most of the assemblies were quite contiguous. This assumption is also reflected by high N50 and NGA50 scores. Interestingly, it was noticed that low NGA50 values correlated with a high number of misassemblies. A similar phenomenon was previously observed by Jünemann et al., however, he concluded that direct concordance between low NGA50 values and high rates of misassemblies has a rather little meaning (Jünemann et al., 2014).

The correctness of the sequence assembly is estimated by examining whether the assembled contigs accurately represent the genome. This can be inspected with use of BLAST search tool against genomes which are already sequenced, or against local databases. By doing so, it was observed that SPAdes integrated plasmidal DNA into chromosomes. SPAdes tendency to join plasmids with chromosomes has been noticed previously by Judge et al., who used SPAdes for assembly of combined MinION and Illumina data (Judge et al., 2016). Although merging plasmids to a chromosome, statistics generated by QUAST for both, SPAdes and A5-pipeline, ranked SPAdes as an overall better assembler.

Regarding plasmids predicted in this study, a majority of these sequences presented homology to chromosomes of various *Campylobacter* strains. The possible reason for this might be that newly sequenced genomes are being uploaded into GenBank without determining of the plasmidal sequences. In fact, many BLAST hits homologous with plasmids from this study (Table 7), were assembled by SPAdes (NCBI, 2018c), and perhaps for these genome assemblies, the plasmidal DNA sequences were integrated into chromosomes.

Moreover, it was marked that plasmids predicted for the genome of *C. jejuni* (isolate 16EP001139) was homologous to *C. coli* plasmid. The opposite situation was observed for *C. coli* isolate 15EP000596, the plasmid of which was homologous with the pTet plasmid of *C. jejuni* strain S3. This phenomenon suggests that transmission of the DNA via horizontal gene transfer can occur between *C. jejuni* and *C. coli.* This observation was reported before (Boer et al., 2002; Wang & Taylor, 1990).

4.2 cgMLST and phylogeny

A cgMLST approach based on 637 loci shared by *Campylobacter* isolates allowed to distinguish three clusters with ≤ 2 allelic differences between the isolates (Figure 14). Cluster A (Table 9) consisted of isolates known to be from an outbreak of campylobacteriosis among children after a visit to a farm in May 2011 (Møller-Stray et al., 2012). The cgMLST approach has designated these isolates as ST-21, which is often reported from farms (Kwan et al., 2008). Moreover, no allelic differences between these isolates were detected, which indicates their close epidemiological link. Similarly, clusters B and C had ≤ 2 allelic differences between the isolates and were described as isolates from previous outbreaks. Interestingly, isolates 17EP001087, 17EP001093, 17EP001096, and 17EP001113 were thought to be an outbreak based on the period of occurrence (Table 3). However, analysis of WGS data has verified that only two of them could be considered as an outbreak (cluster C in Table 9). Remaining two isolates were vastly distant and belonged to different ST types, compared to the outbreak cluster.

Clustering of isolates in the core genome-based phylogeny was in agreement with the clustering by cgMLST (Figure 15). Comparably to the cgMLST method, phylogenetic analysis based on core genome did not detect any nucleotide differences for cluster A. However, within cluster B ≤ 2, and within cluster C 22 nucleotide differences between isolates were identified. On the other hand, SNP-based phylogeny detected ≤ 1 SNP deviations within

cluster A (Figure 17), which seem to suggest that resolution of the SNP approach is higher compared to both, cgMLST and core genome-based phylogeny.

Regarding the choice of phylogeny software, all of them similarly clustered the isolates, with differences observed in branching of the trees. These variations might be caused by divergent implementations of algorithms within the different software, which were used for estimations of relationships between the isolates. PAUP* gave the most reliable results, supported by high bootstrap values, independently of used phylogeny algorithm (supplementary data S8).

Estimations of relationships between analysed isolates were made using different phylogeny algorithms (NJ, ML, MP) in order to find the most accurate tree topology. All these methods resulted in similar outcomes, therefore we believe that the described relationships are very probable. However, there were few differences between the trees constructed using ML and NJ methodology, mostly visible on branches, where both techniques had low bootstrap values. The variations between all three methods are due to different assumptions and algorithms that they are based on (Kuhner & Felsenstein, 1994).

C. jejuni and *C. coli* are phenotypically homogenous (On, 2005), therefore the accurate differentiation between these species might be impeded. Phylogenetic investigation of the isolates based on the core genome and SNP level has revealed that two isolates that had been phenotypically classified as *C. jejuni* were, in fact, *C. coli*. This circumstance, in turn, shows the advantage of WGS-based methods over phenotypic methods for correct identification of *Campylobacter* species.

4.3 Antimicrobial resistance and virulence

Within analysed isolates, a predominance (26/40) of non-wild type *Campylobacter* was observed, based on EUCAST resistance breakpoints. Moreover, a high degree of resistance among isolates was travel-associated. A majority of multiresistant strains was imported to Norway from countries with an increased prevalence of campylobacteriosis infections (EFSA & ECDC, 2017).

The phenotypic resistance was towards CIP (24% for *C. jejuni* and 85% for *C. coli*), TET (21% for *C. jejuni* and 85% for *C. coli*), NX (24% for *C. jejuni* and 85% for *C. coli*), ERM (27% for *C. jejuni* and 71% for *C. coli*) and to a less extent to GEN (12% for *C. jejuni* and 42% for *C. coli*). Bioinformatic analysis of WGS data has shown a high degree of correlation between phenotypic resistance to a specific antibiotic, and the presence of

corresponding resistance genes. However, in some instances, antimicrobial resistance genes were not detected within isolates presenting phenotypic resistance to an antimicrobial agent. Isolates 12EP001377, 13EP000100, 14EP000843, 14EP000843, 16EP000713, and 16EP001980 revealed phenotypic resistance towards ERM without carrying *cmeA, cmeB, cmeC* genes; isolates 12EP000401, 13EP001978, and 14EP001612 with resistance to TET did not carry *tetO* or *tet(W/N/W)* genes; isolates 13EP001161, 14EP001617, and 14EP001642 presented resistance to GEN without carrying APH(2") family genes. A similar observation was noted by Han et al. (Han et al., 2016): in their research 7 out of 130 TET-resistant *Campylobacter*, did not carry the *tetO* gene. These findings may suggest that *Campylobacter* have more complex resistance mechanisms than the known antibiotic efflux pumps, antibiotic target protection, or antibiotic inactivation. A review by Engberg et al. (Engberg et al., 2001) describes that mutations in ribosomes as well as in genes encoding subunits of DNA gyrase are responsible for resistance against macrolides, fluoroquinolones, quinolones, and even tetracycline. Therefore, defining the actual phenotypic resistance using the WGS data demands more targeted research, based on finding specific mutations within a genome.

Interestingly, 30/40 *Campylobacter* isolates carried genes encoding cmeABC efflux pump, however, only 14 presented phenotypic resistance towards ERM or CIP. This might suggest that a decrease in the MIC for macrolide- and fluoroquinolone-resistant isolates was caused by inactivation of the *cmeA*, *cmeB*, *cmeC* resistance genes. Moreover, resistance to ERM was not always correlated with resistance to CIP. This observation, on the other hand, indicates the significance of mutations within ribosomes or/and genes encoding the DNA gyrase, which can be associated with resistance towards macrolides and fluoroquinolones. It is speculated that in this research, some isolates might have carried some target point mutations or another yet-uncharacterized mechanism associated with resistance towards CIP or ERM only.

In addition to genes encoding cmeABC efflux pump, the *tetO* and aminoglycoside resistance genes, the WGS approach distinguished presence of $bla_{oxa\text{-}61}$ gene in the majority of *C. jejuni* and *C. coli* isolates. This gene is associated with resistance to ampicillin and cephalosporins (Alfredson & Korolik, 2005). Ampicillin was not included in this research as it is not used in campylobacteriosis treatment; therefore, the correlation between resistance phenotype and genotype for this antibiotic could not be evaluated. However, it is reported that *Campylobacter* produce β-lactamase to enhance their resistance to this antibiotic (Tajada, Gomez-Graces, Alos, Balas, & Cogollos, 1996).

Considering virulence factors, BLASTn searches enabled to exclude plasmid-encoded virulence genes from the analysed data set. Research has shown that none of the examined isolates has harboured pVir plasmid with the type IV secretion system, responsible for microtubule-dependent invasion pathway ("Virulence factors of pathogenic bacteria," 2018). All virulence genes found during the study were chromosome-encoded and were responsible for essential survival factors such as motility, chemotaxis, and adherence (data not shown). Moreover, no analysed isolates were proven to carry genes encoding CDT.

Although WGS brought many insights into resistance genotype and virulence of *Campylobacter*, some limitations of this technique need to be mentioned. Firstly, the predictions of both, AMR and virulence genes were made based on the databases including genes that have been previously determined. Detection of new genes with use of WGS is therefore impeded. Secondly, the output of sequencing methods involves the fragmentary genomes in the form of contigs or scaffolds, which permit some genes to go undetected. Furthermore, the study has shown some discrepancies between susceptibility and genotype; the WGS-based analysis does not present the information about gene expression levels, and therefore the phenotypic susceptibility testing is necessary to describe resistance.

5. Conclusion

Sequencing of high-throughput data followed by genome assembly demands a selection of methods and software for obtaining reliable outputs. Here, we performed WGS of 40 *Campylobacter* isolates on the Illumina MiSeq platform and assembled the genomes using SPAdes *de novo* assembler and reference-based mapping with BWA. Chosen methods resulted in good quality assemblies which could successfully be used in further analysis.

In this study, the WGS data strengthened the analysis of evolutionary relationships between *Campylobacter* isolates, and enabled to distinguish three outbreak clusters, differing by ≤ 2 variations either in core genome allelic profiles or SNPs. The results suggest that the clustering threshold for epidemiologically linked *Campylobacter* in Norway should be defined within these limits; however, because of a low number of reported epidemiological outbreaks in this country, this data cannot be taken for granted and needs additional data and further research.

Moreover, WGS accurately predicted resistance phenotypes, as a high degree of correlation between genotypic and phenotypic resistance was observed. However, because WGS does not give insights into gene expression levels, few discrepancies between resistance and genotype were observed. A majority of analysed *Campylobacter* was resistant to at least one tested antimicrobial drug, and a total of 9 multiresistant isolates were observed, based on EUCAST resistance breakpoints. Overall, *C. coli* exhibited higher resistance rates and were more frequently multiresistant, compared to *C. jejuni*.

Additionally, analysis of WGS data led to an observation that SNP-based phylogeny provides higher resolution compared to both, cgMLST and core genome-based phylogeny. Furthermore, all constructed phylogenetic trees presented similar topology, independently of the algorithm used. Therefore, the NJ approach could be recommended for quick analysis of evolutionary distances between taxa, as it is computationally fast. However, for the most rigorous trees, exhaustive algorithms, such as ML should be utilized.

WGS shows great potential in surveillance and investigation of epidemiological links. Along with decreasing prices of bacterial genome sequencing and high-resolution clustering methods, this technique can successfully be applied in the routine surveillance and detection of infections caused by *Campylobacter* species.

6. Future perspectives

Along with appearance of third-generation sequencing platforms such as PacBio and Oxford Nanopore, we can expect that more accurate assemblers will be developed, enabling processing of bioinformatic data by a broader spectrum of users. Furthermore, dropping sequencing costs might empower WGS-based methods in routine epidemiological investigations.

For purposes of WGS-based surveillance of campylobacteriosis, international standardization is needed, with regards to local and global comparability of *Campylobacter* infections. The standardization of these techniques has to encounter the genomic diversity of *Campylobacter* populations, as well as the implementation of knowledge about pathogenicity and molecular mechanisms of these species. Once the criteria of campylobacteriosis investigation are established, the WGS-based surveillance techniques can help the public health sector to monitor and share information to prevent and control future outbreaks. Further studies with use of third-generation sequencing might bring meaningful data about *Campylobacter*, especially *C. coli*, and fill up the knowledge gaps regarding the diagnosis and prevention of campylobacteriosis.

7. References

Aarestrup, F. M., & Engberg, J. (2001). Antimicrobial resistance of thermophilic Campylobacter. *Veterinary research, 32*(3-4), 311-321.

Adzitey, F., Huda, N., & Ali, G. R. R. (2013). Molecular techniques for detecting and typing of bacteria, advantages and application to foodborne pathogens isolated from ducks. *3 Biotech, 3*(2), 97-107.

Alfredson, D. A., & Korolik, V. (2005). Isolation and expression of a novel molecular class D β-lactamase, OXA-61, from Campylobacter jejuni. *Antimicrobial agents and chemotherapy, 49*(6), 2515-2518.

Alfredson, D. A., & Korolik, V. (2007). Antibiotic resistance and resistance mechanisms in Campylobacter jejuni and Campylobacter coli. *FEMS Microbiology Letters, 277*(2), 123-132.

Allos, B. M. (1997). Association between Campylobacter infection and Guillain-Barré syndrome. *Journal of Infectious Diseases, 176*(Supplement_2), S125-S128.

Allos, B. M. (2001). Campylobacter jejuni Infections: Update on Emerging Issues and Trends. *Clinical infectious diseases, 32*(8), 1201-1206.

Altekruse, S. F., Stern, N. J., Fields, P. I., & Swerdlow, D. L. (1999). Campylobacter jejuni- an emerging foodborne pathogen. *Emerging infectious diseases, 5*(1), 28.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology, 215*(3), 403-410.

Bae, W., Kaya, K. N., Hancock, D. D., Call, D. R., Park, Y. H., & Besser, T. E. (2005). Prevalence and antimicrobial resistance of thermophilic Campylobacter spp. from cattle farms in Washington State. *Applied and environmental microbiology, 71*(1), 169-174.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology, 19*(5), 455-477.

Bear, R., Rintoul, D., Snyder, B., Smith-Caldas, M., Herren, C., & Horne, E. (2016). *Principles of Biology*. Open Access Textbooks. Book 1. Retrieved from http://newprairiepress.org/textbooks/1

Belén, A., Pavón, I., & Maiden, M. C. (2009). Multilocus Sequence Typing. In *Molecular Epidemiology of Microorganisms* (pp. 129-140): Springer.

Benjamin, A. M., Nichols, M., Burke, T. W., Ginsburg, G. S., & Lucas, J. E. (2014). Comparing reference-based RNA-Seq mapping methods for non-human primate data. *BMC genomics, 15*(1), 570.

Bester, L., & Essack, S. (2008). Prevalence of antibiotic resistance in Campylobacter isolates from commercial poultry suppliers in KwaZulu-Natal, South Africa. *Journal of Antimicrobial Chemotherapy, 62*(6), 1298-1300.

Black, R. E., Levine, M. M., Clements, M. L., Hughes, T. P., & Blaser, M. J. (1988). Experimental Campylobacter jejuni infection in humans. *Journal of Infectious Diseases, 157*(3), 472-479.

Blaser, M. J. (1997). Epidemiologic and Clinical Features of Campylobacter jejuni Infections. *Journal of Infectious Diseases, 176*(Supplement 2), S103-S105.

Blaser, M. J., Wells, J. G., Feldman, R. A., Pollard, R. A., & Allen, J. R. (1983). Campylobacter enteritis in the United States: a multicenter study. *Annals of internal medicine, 98*(3), 360-365.

Boer, P. d., Wagenaar, J. A., Achterberg, R. P., Putten, J. P. v., Schouls, L. M., & Duim, B. (2002). Generation of Campylobacter jejuni genetic diversity in vivo. *Molecular microbiology, 44*(2), 351-359.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120.

Bolton, D. J. (2015). Campylobacter virulence and survival factors. *Food microbiology, 48*, 99. doi:10.1016/j.fm.2014.11.017

Buermans, H., & Den Dunnen, J. (2014). Next generation sequencing technology: advances and applications. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1842*(10), 1932-1941.

Chang, C., & Miller, J. F. (2006). Campylobacter jejuni colonization of mice with limited enteric flora. *Infection and immunity, 74*(9), 5261-5271.

Chaveerach, P., Ter Huurne, A., Lipman, L., & Van Knapen, F. (2003). Survival and resuscitation of ten strains of Campylobacter jejuni and Campylobacter coli under acid conditions. *Applied and environmental microbiology, 69*(1), 711-714.

Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., & Jin, Q. (2005). VFDB: a reference database for bacterial virulence factors. *Nucleic acids research, 33*(suppl_1), D325-D328.

Cody, A. J., Bray, J. E., Jolley, K. A., McCarthy, N. D., & Maiden, M. C. (2017). A core genome multi-locus sequence typing scheme for stable, comparative analyses of Campylobacter jejuni and C. coli human disease isolates. *Journal of Clinical Microbiology*, JCM. 00080-00017.

Cody, A. J., McCarthy, N. M., Wimalarathna, H. L., Colles, F. M., Clark, L., Bowler, I. C., Maiden, M. C., Dingle, K. E. (2012). A longitudinal 6-year study of the molecular epidemiology of clinical Campylobacter isolates in Oxfordshire, United Kingdom. *Journal of Clinical Microbiology, 50*(10), 3193-3201.

Coil, D., Jospin, G., & Darling, A. E. (2014). A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. *Bioinformatics, 31*(4), 587-589.

Davis, L., & DiRita, V. (2008). Growth and laboratory maintenance of Campylobacter jejuni. *Current protocols in microbiology*, 8A. 1.1-8A. 1.7.

Davis, L. M., Kakuda, T., & DiRita, V. J. (2009). A Campylobacter jejuni znuA orthologue is essential for growth in low-zinc environments and chick colonization. *Journal of bacteriology, 191*(5), 1631-1640.

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J. F., Guindon, S., Lefort, V., Lescot, M. (2008). Phylogeny. fr: robust phylogenetic analysis for the non-specialist. *Nucleic acids research, 36*(suppl_2), W465-W469.

Dingle, K., Van Den Braak, N., Colles, F., Price, L. J., Woodward, D. L., Rodgers, F. G., Endtz, H. P., Van Belkum, A., Maiden, M. (2001). Sequence typing confirms that Campylobacter jejuni strains associated with Guillain-Barre and Miller-Fisher syndromes are of diverse genetic lineage, serotype, and flagella type. *Journal of Clinical Microbiology, 39*(9), 3346-3349.

European Food Safety Authority (EFSA). (2018). Campylobacter. Retrieved from https://www.efsa.europa.eu/en/topics/topic/campylobacter

European Food Safety Authority (EFSA) and European Centre for Disease Prevention and Control (ECDC). (2017). *The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2016* (1831-4732).

European Food Safety Authority (EFSA) and European Centre for Disease Prevention and Control (ECDC). (2018). *The European Union summary report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food in 2016*.

Ekblom, R., & Wolf, J. B. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary applications, 7*(9), 1026-1042.

Engberg, J., Aarestrup, F. M., Taylor, D. E., Gerner-Smidt, P., & Nachamkin, I. (2001). Quinolone and macrolide resistance in Campylobacter jejuni and C. coli: resistance mechanisms and trends in human isolates. *Emerging infectious diseases, 7*(1), 24.

Enright, M. C., & Spratt, B. G. (1999). Multilocus sequence typing. *Trends in microbiology, 7*(12), 482-487.

European Committee on Antimicrobial Susceptibility Testing (EUCAST). (2017). EUCAST disk diffusion method. Retrieved from www.eucast.org

Evans, M. R., Roberts, R., Ribeiro, C., Gardner, D., & Kembrey, D. (1996). A milk-borne Campylobacter outbreak following an educational farm visit. *Epidemiology & Infection, 117*(3), 457-462.

Ferrero, R. L., & Lee, A. (1988). Motility of Campylobacter jejuni in a viscous environment: comparison with conventional rod-shaped bacteria. *Microbiology, 134*(1), 53-59.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, JF., Dougherty, B. A., Merrick, J. M. (1995). Whole-Genome Random Sequencing and Assembly of Haemophilus influenzae Rd. *science, 269*(5223), 496-512.

Ge, Z., Schauer, D. B., & Fox, J. G. (2008). In vivo virulence properties of bacterial cytolethal-distending toxin. *Cellular microbiology, 10*(8), 1599-1607.

Geissler, A. L., Bustos Carrillo, F., Swanson, K., Patrick, M. E., Fullerton, K. E., Bennett, C., Barrett, K., Mahon, B. E. (2017). Increasing Campylobacter Infections, Outbreaks, and Antimicrobial Resistance in the United States, 2004–2012. *Clinical infectious diseases, 65*(10), 1624-1631. doi:10.1093/cid/cix624

Gillespie, I. A., O'Brien, S. J., Frost, J. A., Adak, G. K., Horby, P., Swan, A. V., Painter, M. J., Neal, K. R., Campylobacter Sentinel Surveillance System Collaborators. (2002). A Case-Case Comparison of Campylobacter coli and Campylobacter jejuni Infection: A Tool for Generating Hypotheses. *Emerging infectious diseases, 8*(9), 937.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics, 17*(6), 333-351.

Graham, S. W., Olmstead, R. G., & Barrett, S. C. (2002). Rooting phylogenetic trees with distant outgroups: a case study from the commelinoid monocots. *Molecular biology and evolution, 19*(10), 1769-1781.

Griggs, D. J., Peake, L., Johnson, M. M., Ghori, S., Mott, A., & Piddock, L. J. (2009). β-Lactamase-Mediated β-Lactam Resistance in Campylobacter Species: Prevalence of Cj0299 (blaOXA-61) and Evidence for a Novel β-Lactamase in C. jejuni. *Antimicrobial agents and chemotherapy, 53*(8), 3357-3364.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics, 29*(8), 1072-1075.

Han, X., Zhu, D., Lai, H., Zeng, H., Zhou, K., Zou, L., Zou, L., Wu, C., Han, G., Liu, S. (2016). Prevalence, antimicrobial resistance profiling and genetic diversity of Campylobacter jejuni and Campylobacter coli isolated from broilers at slaughter in China. *Food Control, 69*, 160-170.

Harris, S. R., Cartwright, E. J., Török, M. E., Holden, M. T., Brown, N. M., Ogilvy-Stuart, A. L., Ellington, M. J., Quail, M. A., Bentley, S. D., Parkhill, J. (2013). Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. *The Lancet Infectious Diseases, 13*(2), 130-136.

He, Y., Zhang, Z., Peng, X., Wu, F., & Wang, J. (2013). De novo assembly methods for next generation sequencing data. *Tsinghua Science and Technology, 18*(5), 500-514.

Helms, M., Simonsen, J., Olsen, K. E., & Mølbak, K. (2005). Adverse health events associated with antimicrobial drug resistance in Campylobacter species: a registry-based cohort study. *The Journal of infectious diseases, 191*(7), 1050-1055.

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

Huelsenbeck, J. P., & Rannala, B. (1997). Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *science, 276*(5310), 227-232.

Humphrey, T., O'Brien, S., & Madsen, M. (2007). Campylobacters as zoonotic pathogens: a food production perspective. *International journal of food microbiology, 117*(3), 237-257.

Illumina Inc. (2010). Illumina Sequencing Technology. Highest data accuracy, simple workflow, and a broad range of applications. Retrieved from https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Jin, S., Joe, A., Lynett, J., Hani, E. K., Sherman, P., & Chan, V. L. (2001). JlpA, a novel surface-exposed lipoprotein specific to Campylobacter jejuni, mediates adherence to host epithelial cells. *Molecular microbiology, 39*(5), 1225-1236.

Joensen, K., Kuhn, K., Müller, L., Björkman, J., Torpdahl, M., Engberg, J., Holt, H. M., Nielsen, H. L., Petersen, A. M., Ethelberg, S. (2017). Whole-genome sequencing of Campylobacter jejuni isolated from Danish routine human stool samples reveals surprising degree of clustering. *Clinical Microbiology and Infection*.

Jolley, K. A., Chan, M.-S., & Maiden, M. C. (2004). mlstdbNet–distributed multi-locus sequence typing (MLST) databases. *BMC bioinformatics, 5*(1), 86.

Jones, D., Sutcliffe, E., Rios, R., Fox, A., & Curry, A. (1993). Campylobacter jejuni adapts to aerobic metabolism in the environment. *Journal of medical microbiology, 38*(2), 145-150.

Jore, S., Viljugrein, H., Brun, E., Heier, B. T., Borck, B., Ethelberg, S., Hakkinen, M., Kuusi, M., Reiersen, J., Hansson, I., Olddon Engvall, E., Løfdahl, M., Wagenaa, J. A., van Pelt, W., Hofshagen, M. (2010). Trends in Campylobacter incidence in broilers and humans in six European countries, 1997–2007. *Preventive veterinary medicine, 93*(1), 33-41.

Judge, K., Hunt, M., Reuter, S., Tracey, A., Quail, M. A., Parkhill, J., & Peacock, S. J. (2016). Comparison of bacterial genome assembly software for MinION data and their applicability to medical microbiology. *Microbial genomics, 2*(9).

Jünemann, S., Prior, K., Albersmeier, A., Albaum, S., Kalinowski, J., Goesmann, A., Stoye, J., Harmsen, D. (2014). GABenchToB: a genome assembly benchmark tuned on bacteria and benchtop sequencers. *PLoS One, 9*(9), e107014.

KAPABiosystems. (2017). Kapa LTP Library Preparation. In Roche (Ed.).

Kapperud, G., Espeland, G., Wahl, E., Walde, A., Herikstad, H., Gustavsen, S., Tveit, I., Natås, O., Bevanger, L., Digranes, A. (2003). Factors associated with increased and decreased risk of Campylobacter infection: a prospective case-control study in Norway. *American journal of epidemiology, 158*(3), 234-242.

Kapperud, G., Skjerve, E., Bean, N., Ostroff, S., & Lassen, J. (1992). Risk factors for sporadic Campylobacter infections: results of a case-control study in southeastern Norway. *Journal of Clinical Microbiology, 30*(12), 3117-3121.

Kärenlampi, R., Rautelin, H., Schönberg-Norio, D., Paulin, L., & Hänninen, M.-L. (2007). Longitudinal study of Finnish Campylobacter jejuni and C. coli isolates from humans, using multilocus sequence typing, including comparison with epidemiological data and isolates from poultry and cattle. *Applied and environmental microbiology, 73*(1), 148-155.

Koren, S., & Phillippy, A. M. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current opinion in microbiology, 23*, 110-120.

Krosness, M., Lyngstad, T., Lange, H., Nygård, K., Jore, S., Kapperud, G., MacDonald, E., Brandal, L. T., Fergulio, S. L, Grøneng, G. M., Vold, L. (2018). *Overvåkning av infeksjonssykdommer som smitter fra mat, vann og dyr, inkludert vektorbårne sykdommer.* . Retrieved from https://www.fhi.no/publ/2018/overvakning-av-sykdommer-som-smitter-fra-mat-vann-og-dyr/

Kuhner, M. K., & Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular biology and evolution, 11*(3), 459-468.

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular biology and evolution, 35*(6), 1547-1549.

Kwan, P. S., Barrigas, M., Bolton, F. J., French, N. P., Gowland, P., Kemp, R., Leatherbarrow, H., Upton, M., Fox, A. J. (2008). Molecular epidemiology of Campylobacter jejuni populations in dairy cattle, wildlife, and the environment in a farmland area. *Applied and environmental microbiology, 74*(16), 5130-5138.

Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M. R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T. (2015). Insights from 20 years of bacterial genome sequencing. *Functional & integrative genomics, 15*(2), 141-161.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research, 18*(11), 1851-1858.

Lin, J., Akiba, M., Sahin, O., & Zhang, Q. (2005). CmeR functions as a transcriptional repressor for the multidrug efflux pump CmeABC in Campylobacter jejuni. *Antimicrobial agents and chemotherapy, 49*(3), 1067-1075.

Lin, J., Michel, L. O., & Zhang, Q. (2002). CmeABC functions as a multidrug efflux system in Campylobacter jejuni. *Antimicrobial agents and chemotherapy, 46*(7), 2124-2131.

Loman, N. J., Constantinidou, C., Chan, J. Z., Halachev, M., Sergeant, M., Penn, C. W., Robinson, E. R., Pallen, M. J. (2012). High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nature Reviews Microbiology, 10*(9), 599-606.

Loman, N. J., & Pallen, M. J. (2015). Twenty years of bacterial genome sequencing. *Nature Reviews Microbiology, 13*(12), 787-794.

Löytynoja, A. (2014). Phylogeny-aware alignment with PRANK. In *Multiple sequence alignment methods* (pp. 155-170): Springer.

MacDonald, E., White, R., Mexia, R., Bruun, T., Kapperud, G., Lange, H., Nygård, K., Vold, L. (2015). Risk factors for sporadic domestically acquired Campylobacter infections in Norway 2010–2011: A national prospective case-control study. *PLoS One, 10*(10), e0139636.

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the national academy of sciences, 95*(6), 3140-3145.

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

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet., 9*, 387-402.

Mardis, E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature, 470*(7333), 198-203.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature, 437*(7057), 376-380.

McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., Bhullar, K., Canova, M. J., De Pascale, G., Ejim, L. (2013). The comprehensive antibiotic resistance database. *Antimicrobial agents and chemotherapy*, AAC. 00419-00413.

McCarthy, N., & Giesecke, J. (2001). Incidence of Guillain-Barré syndrome following infection with Campylobacter jejuni. *American journal of epidemiology, 153*(6), 610-614.

McPherson, J. D. (2009). Next-generation gap. *Nature methods, 6*, S2-S5.

Meade, K. G., Narciandi, F., Cahalane, S., Reiman, C., Allan, B., & O'Farrelly, C. (2009). Comparative in vivo infection models yield insights on early host immune response to Campylobacter in chickens. *Immunogenetics, 61*(2), 101-110.

Mellmann, A., Bletz, S., Böking, T., Kipp, F., Becker, K., Schultes, A., Prior, K., Harmsen, D. (2016). Real-time genome sequencing of resistant bacteria provides precision infection control in an institutional setting. *Journal of Clinical Microbiology*, JCM. 00790-00716.

Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature Reviews Genetics, 11*(1), 31-46.

Møller-Stray, J., Eriksen, H., Bruheim, T., Kapperud, G., Lindstedt, B., Skeie, Å., Sunde, M., Urdahl, A. M., Øygard, B., Vold, L. (2012). Two outbreaks of diarrhoea in nurseries in Norway after farm visits, April to May 2009. *Eurosurveillance, 17*(47), 20321.

Moore, J. E., Barton, M. D., Blair, I. S., Corcoran, D., Dooley, J. S., Fanning, S., Kempf, I., Lastovica, A., Lowery, C. J., Matsuda, M. (2006). The epidemiology of antibiotic resistance in Campylobacter. *Microbes and infection, 8*(7), 1955-1966.

Nagarajan, N., & Pop, M. (2013). Sequence assembly demystified. *Nature Reviews Genetics, 14*(3), 157-167.

National Human Genome Research Institute. (2016, 6 July 2016). The Cost of Sequencing a Human Genome. Retrieved from https://www.genome.gov/sequencingcosts/

NCBI. (2018a). Campylobacter coli. Genome Assembly and Annotation report. Retrieved 04.09.2018 https://www.ncbi.nlm.nih.gov/genome/genomes/1145

NCBI. (2018b). Campylobacter jejuni Genome Assembly and Annotation report. Retrieved 04.09.2018 https://www.ncbi.nlm.nih.gov/genome/genomes/149

NCBI. (2018c). The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information. Retrieved from https://www.ncbi.nlm.nih.gov

Norwegian Institute of Public Health (NIPH). (2009, 19.11.2014). Utbrudd av campylobacteriose i Norge. Kronologisk oversikt over større utbrudd av campylobacteriose som har vært i Norge siden 1999. Retrieved from https://www.fhi.no/sv/utbrudd/oversikt-over-storre-utbrudd/utbrudd-av-campylobacteriose-i-norg/

Norwegian Institute of Public Health (NIPH). (2018). MSIS- Statistikk. Retrieved from http://msis.no

Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2015). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics, 32*(2), 292-294.

On, S. L. (2005). Taxonomy, Phylogeny, and Methods for the Identification of Campylobacter Species. *Campylobacter: Molecular and Cellular Biology.* Norfolk, United Kingdom: Horizon Press, 13-42.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., Fookes, M., Falush, D., Keane, J. A., Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics, 31*(22), 3691-3693.

Pareek, C. S., Smoczynski, R., & Tretyn, A. (2011). Sequencing technologies and genome sequencing. *Journal of applied genetics, 52*(4), 413-435.

Pattengale, N. D., Alipour, M., Bininda-Emonds, O. R., Moret, B. M., & Stamatakis, A. (2010). How many bootstrap replicates are necessary? *Journal of computational biology, 17*(3), 337-354.

Peterson, M. C. (2003). Campylobacter jejuni enteritis associated with consumption of raw milk. *Journal of Environmental Health, 65*(9), 20.

Pickett, C. L., & Whitehouse, C. A. (1999). The cytolethal distending toxin family. *Trends in microbiology, 7*(7), 292-297.

Poly, F., & Guerry, P. (2008). Pathogenesis of campylobacter. *Current opinion in gastroenterology, 24*(1), 27-31.

Pushkarev, D., Neff, N. F., & Quake, S. R. (2009). Single-molecule sequencing of an individual human genome. *Nature biotechnology, 27*(9), 847-850.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics, 13*(1), 341.

Robinson, D. (1981). Infective dose of Campylobacter jejuni in milk. *British medical journal (Clinical research ed.), 282*(6276), 1584.

Roux, F., Sproston, E., Rotariu, O., MacRae, M., Sheppard, S. K., Bessell, P., Smith-Palmer, A., Cowden, J., Maiden, M. C., Forbes, K. J. (2013). Elucidating the aetiology of human Campylobacter coli infections. *PLoS One, 8*(5), e64504.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution, 4*(4), 406-425.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences, 74*(12), 5463-5467.

Sarovich, D. S., & Price, E. P. (2014). SPANDx: a genomics pipeline for comparative analysis of large haploid whole genome re-sequencing datasets. *BMC research notes, 7*(1), 618.

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. *Human molecular genetics, 19*(R2), R227-R240.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics, 30*(14), 2068-2069.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics, 15*(2), 121.

Smerdon, W., Adak, G., O'Brien, S., Gillespie, I., & Reacher, M. (2001). General outbreaks of infectious intestinal disease linked with red meat, England and Wales, 1992-1999. *Communicable Disease and Public Health, 4*(4), 259-267.

Sougakoff, W., Papadopoulou, B., Nordmann, P., & Courvalin, P. (1987). Nucleotide sequence and distribution of gene tetO encoding tetracycline resistance in Campylobacter coli. *FEMS Microbiology Letters, 44*(1), 153-159.

Strachan, N., Rotariu, O., Smith-Palmer, A., Cowden, J., Sheppard, S., O'Brien, S., Maiden, M. C., Macrae, M., Bessel, P. R., Matthews, L. (2013). Identifying the seasonal origins of human campylobacteriosis. *Epidemiology & Infection, 141*(6), 1267-1275.

Swofford, D. (2003). PAUP*: Phylogenetic analysis using parsimony (* and other methods), version 40b10 Sunderland: Sinauer. In.

Tajada, P., Gomez-Graces, J., Alos, J., Balas, D., & Cogollos, R. (1996). Antimicrobial susceptibilities of Campylobacter jejuni and Campylobacter coli to 12 beta-lactam agents and combinations with beta-lactamase inhibitors. *Antimicrobial agents and chemotherapy, 40*(8), 1924-1925.

Tauxe, R. V., Hargrett-Bean, N., Patton, C., & Wachsmuth, I. (1988). Campylobacter isolates in the United States, 1982-1986. *MMWR CDC Surveill Summ, 37*(2), 1-13.

Thomsen, M. C. F., Ahrenfeldt, J., Cisneros, J. L. B., Jurtz, V., Larsen, M. V., Hasman, H., Aarestrup, F. M., Lund, O. (2016). A bacterial analysis platform: an integrated system for analysing bacterial whole genome sequencing data for clinical diagnostics and surveillance. *PLoS One, 11*(6), e0157718.

Torp, M., Vigerust, M., Bergsjø, B., Er, C., & Hofshagen, M. (2014). The surveillance programme for Campylobacter spp. in broiler flocks in Norway 2015.

Unemo, M., & Dillon, J.-A. R. (2011). Review and International Recommendation of Methods for Typing *Neisseria gonorrhoeae* Isolates and Their Implications for Improved Knowledge of Gonococcal Epidemiology, Treatment, and Biology. *Clinical microbiology reviews, 24*(3), 447-458.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J. A., Costa, G., McKernan, K. (2008). A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome research, 18*(7), 1051-1063.

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in genetics, 30*(9), 418-426.

Virulence factors of pathogenic bacteria. (2018). Retrieved from http://www.mgc.ac.cn/VFs/

Wajid, B., Sohail, M. U., Ekti, A. R., & Serpedin, E. (2016). The A, C, G, and T of Genome Assembly. *BioMed research international, 2016*.

Wang, Y., & Taylor, D. (1990). Natural transformation in Campylobacter species. *Journal of bacteriology, 172*(2), 949-955.

Wieczorek, K., Kania, I., & Osek, J. (2013). Prevalence and Antimicrobial Resistance of Campylobacter spp. Isolated from Poultry Carcasses in Poland. *Journal of Food Protection, 76*(8), 1451-1455. doi:10.4315/0362-028X.JFP-13-035

World Health Organization. (2013). The global view of campylobacteriosis: report of an expert consultation, Utrecht, Netherlands, 9-11 July 2012.

Xiong, J. (2006). *Essential bioinformatics*: Cambridge University Press.

Young, K., Davis, L., & Dirita, V. (2007). Campylobacter jejuni: molecular biology and pathogenesis. *Nature Reviews. Microbiology, 5*(9), 665-679. doi:10.1038/nrmicro1718

Zerbino, D., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, gr. 074492.074107.

Zhao, S., Mukherjee, S., Chen, Y., Li, C., Young, S., Warren, M., Abbott, J., Friedman, S., Kabera, C., Karlsson, M. (2015). Novel gentamicin resistance genes in Campylobacter isolated from humans and retail meats in the USA. *Journal of Antimicrobial Chemotherapy, 70*(5), 1314-1321.

Zhou, J., Zhang, M., Yang, W., Fang, Y., Wang, G., & Hou, F. (2016). A seventeen-year observation of the antimicrobial susceptibility of clinical Campylobacter jejuni and the molecular mechanisms of erythromycin-resistant isolates in Beijing, China. *International Journal of Infectious Diseases, 42*, 28-33.

Supplementary files

1. **S1 – DNA concentrations of isolated *Campylobacter*.pdf**

   Description: The values of DNA concentrations of *Campylobacter* isolated by MagNA Pure 96 system.

2. **S2 – Trimming statistics generated by Trimmomatic.pdf**

   Description: Detailed overview on trimming of Illumina paired-end reads including the size of surviving/dropped nucleotides in bp.

3. **S3 – FastQC quality control reports.pdf**

   Description: FastQC reports for reads before and after trimming process. The reports include i) per base sequence quality, and ii) per base sequence content.

4. **S4 – QUAST statistics for *de novo* assemblies.xlsx**

   Description: Reads assembled by SPAdes and A5-pipeline were assessed by QUAST, which gives insights to various matrices such as number of contigs, N50, and NGA50.

5. **S5 – Predicted plasmids for *Campylobacter* isolates.pdf**

   Description: Overview on BLASTn searches for plasmids predicted by PlasmidSPAdes. The table includes values such as sizes of predicted plasmids, sizes of BLASTn homologs, coverage, and identity.

6. **S6 – Reference-based mapping statistics.xlsx**

   Description: The file includes statistics of reference-based mapping against various reference genomes. Sheet 1 and 2 present statistics for *C. jejuni* and *C. coli*, respectively.

7. **S7 – Results of cgMLST approach.xlsx**

   Description: The file includes the list of the 637 target genes defined by cgMLST approach using Ridom SeqSphere+ (sheet 1), and results of application the developed cgMLST scheme for analysed *Campylobacter* isolates (sheet 2). For the latter, symbol "?" represents the target genes that were not found or failed during the process for the analysed genome.

8. **S8 – Phylogenetic trees generated by CLC and MEGA.pdf**

   Description: The document includes the NJ and ML trees generated by CLC and MEGA for core genome-based phylogeny.

9. **S9 – SNPs summary for single isolates generated by SPANDx.pdf**

   Description: Summary statistics of SNPs for each isolate. These data were obtained after aligning the dataset against *C. jejuni* NCTC 11168 as a reference genome.

10. **S10 –List of antimicrobial resistance genes detected for analysed *Campylobacter* by WGS-based approach.pdf**

    Description: Table presenting antimicrobial resistance genes detected for each isolate by both, BAP and RGI searches against CARD database.