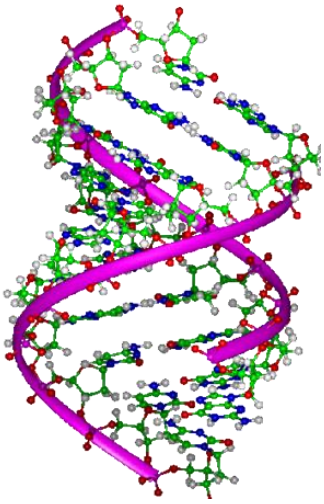
Pangenome-guided Tools for Investigating the Role of Epsilonproteobacteria in Human Gastroenteritis

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

in Veterinary Science

at Massey University, Manawatu, New Zealand



Angela Joyce Cornelius

2018

Abstract

Gastroenteritis affects billions of people every year and current testing methods fail to identify the cause for approximately half of the samples submitted for microbiological testing. Epsilonproteobacteria contains *Campylobacter jejuni*, the most commonly reported cause of bacterial gastroenteritis in the world, and *Helicobacter pylori*, a gastric pathogen and class I carcinogen. This bacterial class also contains ≥20 additional species known, or suspected, of being human pathogens. To better understand the role some of these species play in human gastroenteritis, novel rapid, cost effective methods are needed. The growing number of whole genome sequences available for this class were exploited to first evaluate the classification of the genetically heterogeneous species *C. concisus* and then to identify taxon-specific CDS for a range of Epsilonproteobacterial taxa. Probes were designed to detect 28 of these CDS and incorporated into a single multiplex ligation-dependent probe amplification (MLPA) assay which was tested against DNA from 43 Epsilonproteobacterial species and then applied to DNA extracts from stool samples from a childhood gastroenteritis case control study undertaken in Belgium. The 22 *C. concisus* genomes consistently clustered into two genomospecies (GS) represented by ATCC 33237^T (GS1) and CCUG 19995 (GS2). Taxon-specific genes were identified for 28 taxa, including the two *C. concisus* genomospecies, and concordant results were observed for the majority of MLPA probes and DNA extracts from pure cultures. The probes designed to detect *C. lari* subsp. *concheus* and *H. pullorum* failed to detect the target DNA; all of the urease positive thermophilic *Campylobacter* DNA extracts were also positive for the probe designed to detect *C. subantarcticus*, some probes lacked repeatability in the presence of elevated EDTA and the size differences between some probes needs to be optimised. *C. jejuni* was the most common Epsilonproteobacterial species isolated by culture and *C. concisus* was the most common species detected by MLPA. Both *C. jejuni* and *C. concisus* GS2 were detected in significantly higher numbers in cases than controls in a Belgian

childhood case control study. This demonstrated the utility of the Epsilonproteobacteria MLPA assay and provides some evidence that *C. concisus* GS2 may have a role in childhood gastroenteritis.

Acknowledgements

My first thank you must go to the Institute of Environmental Science and Research (ESR) that has not only provided me with employment for the past 24 years but has also invested in my continued education. To Wim Nijhof, my sincere thanks for your support and encouragement, especially in the last few months, to help get this over the line. Mrs Susan Lin performed the majority of the New Zealand-based laboratory analyses in this project and was always willing to go that extra mile to get the work completed, and to a high standard. She also challenged me, when necessary, to ensure my methods were robust. Dr Darren Smalley and Ms Maud Fazzari also provided exceptional laboratory support and Ms Marilyn Piercy provided expert analyses involving capillary electrophoresis and 16S rRNA sequencing. Dr David Wood demonstrated his exceptional patience helping with R and answering my statistics questions. To Dr Joanne Kingsbury and Dr Lucia Rivas, thank you for the regular coffee excursions that reminded me that there was a world beyond my PhD and gave me new perspectives on problems I was encountering. And thank you to my many colleagues that listened patiently to my latest trials and tribulations or carried the load for me while my focus was elsewhere.

A big thank you goes out to Dr Jan Schouten and Dr Paul van Vught from MRC-Holland who provided advice and support in the MLPA development as well as providing the probes at no cost. Many thanks to Dr Krunoslav Bojanic, Dr Maria Figueras, Dr Mohsina Huq, Dr Jacqueline Keenan, Dr William Miller, Dr Francis Megraud, Professor Olivier Vandenberg and Dr Anne Midwinter for sharing strains, DNA and/or genomes used in the development or testing of the MLPA assay. Thank you, also, to Dr Anne Tillmanne and Professor Olivier Vandenberg for allowing me to be involved in the Belgian childhood gastroenteritis case control study and to Dr Marie Hallin and Dr Delphine Martiny for organising the DNA extractions and shipments to New Zealand.

I gratefully acknowledge the support and sense of community provided by the Computational Biology Team lead by Dr Anthony Poole and Dr Paul Gardner at Canterbury University. Two members require special mention: Nicole Wheeler for agreeing to run my data through her delta bitscore and Fatemeh Ashari Ghomi for walking me through the development of a Python script to help make sense of my BLAST results.

I was exceptionally lucky to have had a supervisory team that included four highly respected scientists with complementary specialist areas. Dr Patrick Biggs introduced me to the wonder of bioinformatics, set up my workstation and guided me so patiently through the world of Linux. Dr Stephen On showed faith in me and my abilities from the beginning and shared his passion for taxonomy. Professor Olivier Vandenberg, provided a clinical microbiology perspective and enthusiasm for ongoing collaborations have been invaluable. Professor Nigel French was patient and supportive as I navigated the world of epidemiology.

Last, but certainly not least, a big thank you to my husband, Malcolm, and son, Ryan, who supported me through this journey even when it limited our time together.

Approval was obtained for the case control study under the project entitled “Detection of emerging campylobacter in children with gastroenteritis” (B076201524271), and the associated modification regarding analysis of part of the sample overseas, from the Ethics Committee of the Centre Hospitalier Universitaire Saint-Pierre (Licence No O.M.007) on 8th April 2015.

The analysis of DNA extracts from the stool samples and isolates associated with the case control study by ESR was lodged with the Massey University Human Ethics Committee in August 2015. This project has been evaluated by peer review and judged to be low risk. Consequently, it has not been reviewed by one of the University’s Human Ethics Committee. Angela Cornelius was responsible for the ethical conduct of this research.

Glossary of Terms

Abbreviations

ΔT_m	Difference in melting temperature
%	Percentage
°C	Degrees Celsius
µg	Micrograms (10^{-6} gram)
µL	Microlitres (10^{-6} litre)
µM	Micromolar (10^{-6} moles/L)
µm	Micrometre (10^{-6} metres)
16S rRNA	Small subunit of ribosomes
23S rRNA	Part of the large subunit of ribosomes
ABI	Applied Biosystems
AFLP	Amplified fragment length polymorphism
AG	Acute gastroenteritis
ag	Attogram (10^{-18} grams)
ANI	Average nucleotide identity
ANIb	Average nucleotide identity using BLAST
ANIm	Average nucleotide identity using MUMmer
ATCC	American Type Culture Collection
BLAST	Basic local alignment search tool
bp	Base pair, nucleotide
BSA	Bovine serum albumin
CBA	Columbia horse blood agar
Ccon	*Campylobacter concisus*
CCUG	Culture Collection, University of Göteborg, Sweden

Ccur	*Campylobacter curvus*
CDS	Coding DNA sequences
CECT	Spanish type culture collection, Universitat de València
Cfet	*Campylobacter fetus*
Chom	*Campylobacter hominis*
CI	Confidence interval
CIP	Collection of the Institut Pasteur, France
Cjej	*Campylobacter jejuni*
Cmuc	*Campylobacter mucosalis*
CO_2	Carbon dioxide gas
COGs	Clusters of orthologous genes
Crec	*Campylobacter rectus*
CSC	Christchurch Science Centre
Csho	*Campylobacter showae*
Cspu	*Campylobacter sputorum*
DDBJ	DNA Databank of Japan
DGGE	Denaturing gradient gel electrophoresis
dH_2O	Deionised water
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
DSM	German collection of microorganisms and cell cultures
EDTA	Ethylendinitrilo tetraacetic acid
EMBL-EBI	European Molecular Biology Laboratory – European Bioinformatics Institute
ER	Emergency room
ESR	Institute of Environmental Science and Research
FAM	6-carboxyfluorescein
FDR	False detection rate

FFP	Feature frequency profiling
fg	Femtogram (10^{-15} grams)
g	Gram
GBDP	Genome BLAST distance phylogeny
GC	Guanine (G) plus cytosine (C) (nitrogenous bases)
GGDC	Genome to genome distance calculator
GroEL	60-kDa chaperonin
GS	Genomospecies
h	hour
H_2	Hydrogen gas
HSPs	High-scoring segment pairs
IPGM	Ion personal genome machine
INSDC	International Nucleotide Sequence Database Collaboration
KCl	Potassium chloride
L	Litre
LAMP	Loop-mediated isothermal amplification
LMG	Laboratory for Microbiology of the Faculty of Sciences at Ghent University, part of the Belgian Coordinated Collections of Microorganisms (BCCM)
LS-BSR	Large scale BLAST score ratio
M	Moles/L
M13	A virus that infects the bacterium *Escherichia coli*
MALDI-TOF	Matrix assisted laser desorption ionisation time-of-flight
Mb	Megabase
MBiT	MLPA-binary typing
mL	Millilitre
MLPA	Multiplex ligation-dependent probe amplification
MLST	Multi-locus sequence typing

min	Minutes
mM	Millimolar (10^{-3} moles/L)
mol%	Molar percentage
MU	Massey University
MUM	Maximal unique matching
N_2	Nitrogen gas
NCBI	National Center for Biotechnology Information (USA)
NCTC	National collection of type cultures, one of the four culture collections operated by Public Health England
ng	Nanogram (10^{-9} gram)
NGS	Next generation sequencing
Ns	Nucleotides or unknown identity (could be adenine [A], cytosine [C], guanine [G] or thymine [T])
NZ	New Zealand
NZGL	New Zealand Genomics Ltd
NZRM	New Zealand Reference Culture Collection, Medical Section
O_2	Oxygen gas
OR	Odds ratio
PCR	Polymerase chain reaction
PFGE	Pulsed-field gel electrophoresis
pg	Picogram (10^{-12} grams)
RFLP	Restriction fragment length polymorphism
RMIT	RMIT university, Melbourne, Australia
rMLST	Ribosomal multi-locus sequence typing
ROC	Receiver operator characteristics
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid

^{T}	Type strain
TBE	Tris-borate-EDTA buffer
TE	Tris-EDTA buffer
Tetra	Tetranucleotide frequency
TG-ROC	Two-graph receiver operator characteristic analysis
T_m	Melting temperature
tmRNA	Transfer messenger ribonucleic acid
tRNA	Transfer ribonucleic acid
Tris	Tris(hydroxymethyl)aminomethane
TSA	Tryptic soy agar
U	Units
UPTC	Urease-positive thermophilic *Campylobacter*
USA	United States of America
USDA	United States Department of Agriculture
V	volts
wgMLST	Whole genome MLST
WGS	Whole genome sequences or whole genome sequencing

Multiplex Ligation-dependent Probe Amplification (MLPA) Probes

Abutzleri	MLPA probes designed to detect *Arcobacter butzleri*
Acryaerophilus	MLPA probe designed to detect *Arcobacter cryaerophilus*
Arcobacter	MLPA probes designed to detect all species within the *Arcobacter* genus
Campylobacter	MLPA probes designed to detect all of the species with the *Campylobacter* genus
Cavium	MLPA probe designed to detect *Campylobacter avium*
Ccanadensis	MLPA probe designed to detect *Campylobacter canadensis*
Ccoli	MLPA probes designed to detect *Campylobacter coli*

Cconcisus	MLPA probes designed to detect *Campylobacter concisus*
CconcisusGS1	MLPA probes designed to detect *Campylobacter concisus* GS1
CconcisusGS2	MLPA probes designed to detect *Campylobacter concisus* GS2
Ccuniculorum	MLPA probe designed to detect *Campylobacter cuniculorum*
Chelveticus	MLPA probe designed to detect *Campylobacter helveticus*
Cinsulaenigrae	MLPA probe designed to detect *Campylobacter insulaenigrae*
Cjejuni	MLPA probes designed to detect *Campylobacter jejuni*
Cjejunidoylei	MLPA probe designed to detect *Campylobacter jejuni* subsp. *doylei*
Clariconcheus	MLPA probe designed to detect *Campylobacter lari* subsp. *concheus*
Clarilari	MLPA probe designed to detect *Campylobacter lari* subsp. *lari*
ClariUPTC	MLPA probe designed to detect UPTC
Cpeloridis	MLPA probe designed to detect *Campylobacter peloridis*
Csubantarcticus	MLPA probe designed to detect *Campylobacter subantarcticus*
Cupsaliensis	MLPA probe designed to detect *Campylobacter upsaliensis*
Cureolyticus	MLPA probe designed to detect *Campylobacter ureolyticus*
Cvolucris	MLPA probe designed to detect *Campylobacter volucris*
eHelicobacter	MLPA probe design to detect the *Helicobacter* species associated with the intestine of humans and animals (*H. bilis, H. canis, H. cinaedi, H. fennelliae, H. hepaticus, H. pametensis* and *H. pullorum*)
Hcanis	MLPA probe designed to detect *Helicobacter canis*
Hcinaedi	MLPA probe designed to detect *Helicobacter cinaedi*
Hfennelliae	MLPA probe designed to detect *Helicobacter fennelliae*
Hpullorum	MLPA probe designed to detect *Helicobacter pullorum*

Other Terms

blastn	BLAST search of nucleotide database using a nucleotide query
blastp	BLAST search of protein database using a protein query
blastx	BLAST search of protein database using a translated nucleotide query
BLAST+	Updated version of the BLAST suite available from the NCBI
BSR	BLAST score ratio, the BLAST score for a query peptide divided by the self-BLAST score of the reference peptide
contig	A set of overlapping DNA segments that together represent a consensus region of DNA
*cpn*60	Gene encoding the 60-kDa chaperonin (also known as *gro*EL and *hsp*60)
csv	Comma-separated value, a type of file format
DBS	Delta-bitscore, a profile-based homology scoring method
DRYAD	A curated general-purpose repository of data underlying scientific and medical publications that makes data discoverable, freely reusable, and citable
EpsiloFaa	A local BLAST database containing amino acid sequences from 939 Epsilonproteobacteria genomes
EpsiloFsa	A local BLAST database containing DNA sequences from 939 Epsilonproteobacteria genomes
Excel	A spreadsheet application developed by Microsoft
GenBank	An open access, annotated collection of all publicly available nucleotide sequences and their protein translations that is produced and maintained by NCBI as part of the INSDC
Geneious	A comprehensive suite of molecular biology and NGS analysis tools developed by the company Biomatters

Grade	A percentage calculated by Geneious by combining the query coverage, e-value and identity with weights of 0.5, 0.25 and 0.25, respectively
*gro*EL	Gene encoding the 60-kDa chaperonin (also known as *cpn*60 and *hsp*60)
*gyr*B	Gene encoding the DNA gyrase subunit B
*hae*IIM	Gene encoding a modification methylase
Heatmap	A graphical representation of data where the individual values contained in a matrix are represented as colours
*hip*O	Gene encoding the N-benzoylglycine amidohydrolase (hippuricase) enzyme; annotated as *yxe*P by Prokka
*hsp*60	Gene encoding the 60-kDa chaperonin (also known as *gro*EL and *cpn*60)
kmer	All the possible substrings of length k that are contained in a string, called *l*-mer by Feature Frequency Profiling
KrunoFsa	A local BLAST database containing DNA sequences from three *C. helveticus* and three *C. upsaliensis* genomes shared by Dr Krunoslav Bojanic
l-mer	All the possible substrings of length l that are contained in a string (also called kmer)
LHS	Left hybridization sequence, the portion of the LPO than anneals to target the DNA sequence
locus_tag	Identifiers that are systematically applied to every gene in a genome within the context of sequencing projects
LPO	Left probe oligonucleotide, half of the MLPA probe used to detect DNA sequences
*lpx*A	Gene encoding UDP-N-acetylglucosamine acetyltransferase
mfold	Software that calculates the minimum free energy (ΔG) for DNA and RNA folding
MUMmer	An ultra-rapid alignment tool that uses maximal unique matching

N_{50}	A measure of the quality of a whole genome assembly (50% of the assembled bases are in a contig of this size or larger)
NeighborNet	An algorithm for constructing phylogenetic networks which is loosely based on the neighbor joining algorithm
NEXUS	An extendable file format for storing systematic data for use by computer programs
OrthoMCL	Ortholog group identification using the Markov cluster algorithm
Perl	A computer programming language
Pfam	A database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models
phmmer	A protein search algorithm that uses hidden Markov models
Prokka	A rapid prokaryotic genome annotation software program
pubMLST	A curated online database of MLST data
Q-fragments	Small oligonucleotides bound by PCR primer sequences that are included in MLPA assays as indicators of low target DNA quantities
QUAST	A quality assessment tool for genome assemblies
R	A computer language and environment for statistical computing
RawProbe	A program for calculating the melting temperature of a DNA sequence
RHS	Right hybridization sequence, the portion of the RPO than anneals to target the DNA sequence
RPO	Right probe oligonucleotide, half of the MLPA probe used to detect DNA sequences
RMITConcisus	A local BLAST database containing DNA sequences from four *C. concisus* genomes shared by Dr Mohsina Huq
Roary	A pan-genomic analysis tool

SP	Spanning probe, an optional third component of a MLPA probe that anneals to the target DNA sequence between the LHS and RHS and necessitates a second ligation event
SplitsTree4	An interactive and comprehensive tool for inferring phylogenetic networks from sequences, distances and trees
Stuffer	A small non-annealing DNA sequence used to extend the length of MLPA probes
tblastn	BLAST(+) search of translated nucleotide database using a protein query
tblastx	BLAST(+) search of translated nucleotide database using a translated nucleotide query
*yxe*P	Gene encoding a putative hydrolase YxeP (also known as *hip*O)

Table of Contents

List of Tables

List of Figures

Chapter 1: General Introduction

1.1 Background

Every year billions of people (GBD 2015 Disease and Injury Incidence and Prevalence Collaborators 2016), especially children (King et al. 2003), experience gastroenteritis and the cause is often not identified (Lake et al. 2009). *Campylobacter jejuni* and *C. coli* are collectively the most commonly reported bacterial cause of gastroenteritis in the world (Anonymous 2012). The bacterial class Epsilonproteobacteria also contains several other species known to cause gastroenteritis and a larger number of species that, although associated with the illness, have not been proven to have a causative role (Engberg et al. 2000). Current methods employed by most clinical microbiology laboratories rely on culture for the detection of bacterial pathogens which are time consuming and laborious (Zhang, Morrison, and Tang 2015) and, generally recover only *C. jejuni* and *C. coli* from the class Epsilonproteobacteria (Lastovica 2006). No attempt is generally made to isolate or detect other known, or suspected pathogenic species from *Campylobacter, Arcobacter* or *Helicobacter*. A growing number of commercially available genetic tests are available for the simultaneous detection of bacteria, viruses and protozoa with proven roles in gastroenteritis but they, too, generally only include *C. jejuni* (Zhang, Morrison, and Tang 2015) from the *Campylobacter* genus and no published method simultaneously detects taxa from *Arcobacter, Campylobacter* and *Helicobacter*. More rapid methods are required to provide actionable results for patient care and to control the transmission of pathogens in outbreak settings. Methods for the detection of a broader range of Epsilonproteobacterial taxa are required in order to better understand their roles in the aetiology of human gastroenteritis.

With the recent developments in DNA sequencing technologies, increasing numbers of bacterial whole genome sequences (WGS) have been generated (Chun and Rainey 2014), including thousands of genomes representing species within the Epsilonproteobacteria.

Alongside these sequencing technologies has been the development of bioinformatics tools for comparing WGS. These tools have increased in complexity to deal with the larger datasets that are available for analysis.

WGS were generated for eight well characterised, genetically diverse *C. concisus* isolates to complement genomes available from NCBI (n = 9) and collaborators (n = 5). Eight bioinformatics tools were used on this collection of 22 *C. concisus* genomes, along with nine genomes representing related species, to provide a better understanding of the classification of *C. concisus*.

Additional bioinformatic tools were applied to 939 Epsilonproteobacteria WGS in six separate analyses (Epsilonproteobacteria [939 WGS], *Arcobacter* [21 WGS], *Campylobacter* [329 WGS], *Helicobacter* [428 WGS], *C. concisus* [17 WGS] and *C. lari* group [6 WGS]) which facilitated the identification of taxon-specific coding sequences (CDS) for 27 taxa. A profile-based homology scoring method and published PCR methods were used to identify suitable genes for the detection of five additional taxa. Multiplex ligation-dependent probe amplification (MLPA) probes were designed for 28 of these taxon-specific CDS/genes and then manufactured, and combined to generate probemixes, by the Dutch biotechnology company MRC-Holland. The resulting Epsilonproteobacteria MLPA assay was tested against 127 DNA extracts representing 80 taxa and receiver operator characteristic analysis was conducted to provide a measure of the success of the assay at identifying the target taxa.

This assay was then applied to a collection of 337 DNA extracts from stool samples associated with a case control study conducted in Belgium involving children presenting at the emergency room of two hospitals with gastroenteritis and age-matched controls without signs of gastroenteritis that were attending general clinics at the same hospitals. The culture and MLPA results for the Epsilonproteobacterial class were compared to evaluate how suitable MLPA was for microbiological analysis of stool samples. The statistical significance of both culture and MLPA results were evaluated using both univariate and multivariate analysis to provide a better

understand of the roles a broader range of Epsilonproteobacterial taxa have in human gastroenteritis.

1.2 Research Questions

There were four main research questions in this thesis:

1. What can WGS-based comparative analysis tools tell us about the classification of the genetically heterogeneous species *C. concisus*?
2. Can the increasing number of WGS be exploited to identify taxon-specific genes suitable for incorporation into an Epsilonproteobacteria MLPA assay?
3. Can MLPA provide sensitive detection of a range of taxa from complex matrices such as human stool samples?
4. Are Epsilonproteobacteria taxa other than *C. jejuni* and *C. coli* risk factors for the development of childhood gastroenteritis?

1.3 Structure of this Thesis

The aim of this thesis was to use whole genome sequence data, and associated bioinformatics tools, to first evaluate the classification of a genetically heterogeneous species *C. concisus* and then to design a new molecular assay for the detection of a range of 28 Epsilonproteobacterial taxa. This assay would then be applied to samples from a childhood gastroenteritis case control study involving 184 cases and 177 controls in order to evaluate whether there was associations between Epsilonproteobacterial taxa and this illness. The thesis was structured around the research questions.

Chapter 2 is a literature review covering the burden and aetiology of gastroenteritis; taxonomy and the process for describing new taxa; a description of the taxa within the bacterial class Epsilonproteobacteria; and the laboratory methods routinely used and published for the detection of Epsilonproteobacterial taxa. It also describes the advancements within DNA

sequencing and bioinformatics that facilitate the development of new tools for detection of bacterial taxa.

Chapter 3 is a genomic analysis of the genetically heterogeneous species *C. concisus* that has been isolated from both healthy and diarrhoeic stool samples. A collection of 22 *C. concisus* genomes, and nine genomes representing related species, were compared using a range of bioinformatics tools with the aim of better understanding the classification of this species.

Chapter 4 describes the development of a multiplex ligation-dependent probe amplification (MLPA) assay for the simultaneous detection of 28 taxa within Epsilonproteobacteria. Bioinformatics tools were used to identify coding sequences (CDS) suitable for taxon-specific detection and the probemixes were generated by the Dutch biotechnology company MRC-Holland. The assay was then tested against a collection of control DNA covering a broad range of Epsilonproteobacterial taxa and a small selection of bacterial pathogens associated with gastroenteritis.

In Chapter 5, the MLPA assay was applied to DNA extracted from stool samples collected as part of a Belgian gastroenteritis childhood case control study. The aim of this study was to evaluate the application of a MLPA assay to DNA extracted directly from human faecal samples and to evaluate whether there are any associations between Epsilonproteobacterial taxa and acute gastroenteritis.

This thesis concludes with Chapter 6 which includes a discussion of the findings from these studies and the implications for Epsilonproteobacterial taxonomy, pathogen detection and the epidemiology of childhood gastroenteritis. The limitations of the current study along with future directions are also discussed.

Four appendices and three supplementary files accompany this thesis. Appendix I includes the NeighborNets, generated using a range of *l*-mer lengths, for feature frequency profiling of 31 genomes representing *C. concisus* and related species. Appendix II and Supplementary File 1 contain the ribosomal multi-locus sequence typing alleles and number of

genes identified for the functional groups from the clusters of orthologous genes databases, respectively, for the same set of 31 genomes. Supplementary File 2 is a summary of the 949 Epsilonproteobacterial genomes used in this study and Appendix III details the generation and 16S rRNA sequence confirmation of the identification of control DNA used to test the MLPA assay. Appendix IV contains the Epsilonproteobacteria MLPA assay specificity results and Supplementary File 3 summarises the relevant information form the Belgian childhood gastroenteritis case control study.

Chapter 2: Literature Review

2.1 Abstract

Over two billion cases of acute gastroenteritis occur globally each year with a causative agent being identified in only a proportion of samples submitted for microbiological analysis. The bacterial class Epsilonproteobacteria contains *Campylobacter jejuni* and *C. coli*, which are collectively the most commonly reported cause of human bacterial gastroenteritis worldwide, and *Helicobacter pylori*, the first bacterial species classified as a carcinogen due to its role in gastric cancer. In addition, at least 20 other species, in four genera, have been associated with gastroenteritis, although the causative role has not yet been established for some of these species. These Epsilonproteobacterial species have diverse growth requirements and are often difficult to identify using traditional methods meaning few laboratories routinely test human stool samples for species other than *C. jejuni* and *C. coli*.

To better understand the possible causes of human gastroenteritis, new methods for detecting a broad range of species in human stool samples need to be developed. In this review, the literature concerning the burden and causes of human gastroenteritis will be summarised. The taxonomy of Epsilonproteobacteria, and the current consensus on the roles various taxa in this class play in human gastroenteritis, will also be reviewed. Culture, antigenic and genetic methods for the detection of Epsilonproteobacterial taxa in human stool samples will be summarised. Finally, advancements in whole genome sequencing and the bioinformatics tools that have been developed to facilitate core and pan-genomic analysis will be reviewed.

2.2 Gastroenteritis

Gastroenteritis is defined as an inflammation of the stomach and intestine (Anonymous 1949) with symptoms typically including diarrhoea, stomach cramps, nausea and vomiting (Adlam et al. 2011). An estimated 2.39 billion cases of diarrhoeal disease occurred globally in

2015 (GBD 2015 Disease and Injury Incidence and Prevalence Collaborators 2016), and approximately 4.6 million episodes of gastroenteritis occur each year in New Zealand, averaging 1.11 episodes/person per year (Adlam et al. 2011). For each notified case in New Zealand approximately 13 stool samples are tested, 49 cases visit a General Practitioner and 220 people are sick in the community (Figure 1) (Lake et al. 2010). These results are similar to those observed in Canada (Lake et al. 2010) however the number of community cases per notified case were almost double the 136 cases observed for the United Kingdom (Wheeler et al. 1999). For the majority of cases the illness results in loss of time at work, school or recreation and, from a national perspective, approximately 4.5 million days of paid work are lost to acute gastroenteritis in New Zealand each year (Adlam et al. 2011). When the personal, industrial and governmental costs of foodborne gastroenteritis are combined, the estimated annual cost to New Zealand is NZ$161.9 million (Gadiel and Abelson 2010). Globally, an estimated 549 million cases of foodborne gastroenteritis occurred in 2010, with 349 million of these cases being attributable to bacteria (Foodborne Disease Burden Epidemiology Reference Group 2007-2015 2015).

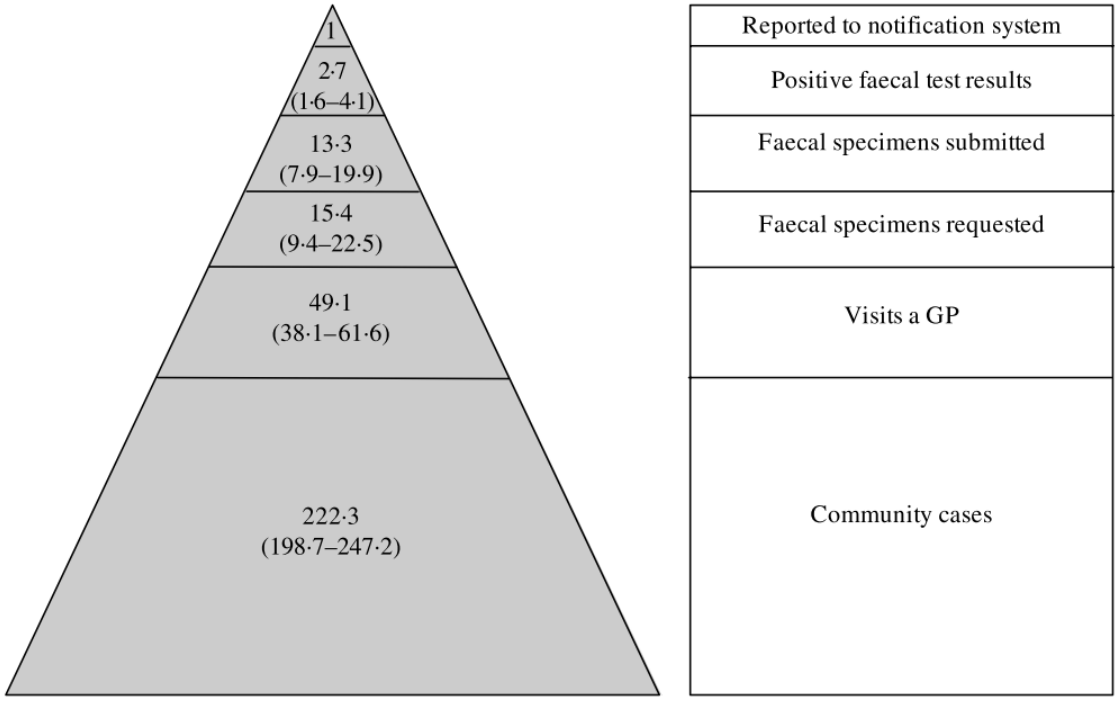


Figure 1: New Zealand Acute Gastroenteritis Reporting Pyramid, 2006
Shows the ratio of cases in the community, general practice, and clinical laboratory levels relative to notifiable diseases (mean, 5th and 95th percentiles). Reproduced from Lake et al. 2010. Epidemiol. Infect. 138:1468

Children under five years have the highest incidence of acute gastroenteritis with an estimated 1.5 billion cases and 1.5-2.5 million deaths globally per year (King et al. 2003). This age group also has the highest prevalence of acute gastroenteritis in New Zealand (Adlam et al. 2011).

There is no consensus on the causes of acute gastroenteritis. Commonly reported causes include the bacteria *Salmonella* spp., *Shigella* spp., *Campylobacter* spp., *Clostridium difficile*, and pathogenic *Escherichia coli*; the viruses rotavirus, caliciviruses (including norovirus), adenovirus types 40 and 41 and astrovirus; and the protozoa *Giardia lamblia, Cryptosporidium* spp. and *Entamoeba histolytica* (Barrett and Brown 2016, Blacklow and Greenberg 1991, Elliott 2007, Guerrant and Bobak 1991, Musher and Musher 2004).

A pathogen was identified in only 23.5% of the estimated 250,000 stool samples submitted for microbiological analysis in New Zealand during 2005 (Lake et al. 2009). In a more

recent New Zealand study, a pathogen was detected by conventional testing (culture, microscopy, immunoassay and viral PCR) in 18% of samples but the proportion of samples positive for a pathogen increased to 30% when a broad range multiplex PCR approach was used (McAuliffe et al. 2013). Reported rates of pathogen detection from other countries varies from 6.4% (de Boer et al. 2010) to 98% (Friesema et al. 2012), although these more extreme values should be considered as outliers as most reported rates were between 20% and 75% (Amar et al. 2007, Boga et al. 2004, Bresee et al. 2012, Cheun et al. 2010, Colomba et al. 2006, Coupland et al. 2013, de Boer et al. 2010, de Wit, Koopmans, Kortbeek, van Leeuwen, Bartelds, et al. 2001, de Wit, Koopmans, Kortbeek, Wannet, et al. 2001, de Wit, Koopmans, Kortbeek, van Leeuwen, Vinje, et al. 2001, Fiedoruk et al. 2015, Klein et al. 2006, Lausch et al. 2017, Lorrot et al. 2011, Olesen et al. 2005, Tam et al. 2012, Tompkins et al. 1999).

The cause of the unsolved cases of disease, including gastroenteritis, has been of great scientific interest for many years and guidelines for establishing causation have been available for over 120 years. Koch's postulates, published in 1891, stipulate that in order to infer that an agent causes a disease it must be present in every case of the disease, it must be specific for the disease, and once isolated in pure culture it can reproduce the disease in a naïve host (Evans 1976). Since this time there have been several alternative strategies published for inferring causation, in line with technological advances (Evans 1976, Firth and Lipkin 2013), and culminating in the staged and confidence approaches to causation (Lipkin 2010). These two approaches to inferring causation are summarised and compared to Koch's postulates in Table 1.

Table 1: Strategies for Inferring Causation

Koch's Postulates (Evans 1976)	Staged Approach to Causation (Lipkin 2010)	Confidence Approach to Causation (Lipkin 2010)
1. The agent is present in every case of the disease 2. The agent is not found in association with any other disease 3. After being isolated in pure culture, the agent can induce the disease in a naïve host	1. Detect an agent or its footprints in association with disease 2. Provide a plausible mechanism for an explanation of disease 3. Demonstrate that modulation of the agent concentration, or of a factor that can be attributed to the presence of the agent (e.g. an antibody), influences the presence or severity of disease 4. Demonstrate that preventing infection prevents disease	1. Possible causal relationship - Statistical association between an agent and a disease - Precedent indicating biological plausibility increases confidence 2. Probable causal relationship - Interventional evidence - Prophylactic evidence 3. Definitive causal relationship - Fulfilled Koch's postulates or a variant

Several epidemiological study designs provide information on the associations between disease and risk factors. The cross-sectional study collects information about the current disease state and/or current exposure status and so measures prevalence (Silman and Macfarlane 2002, 35-37). The case control study compares the exposures to risk factors of individuals with the disease (cases) and individuals without the disease (controls) (Silman and Macfarlane 2002, 37-39). The cohort study involves following one or more groups of individuals, defined by their exposure status, through time to identify disease onset with the aim of determining whether initial exposure status influences the risk of subsequent disease (Silman and Macfarlane 2002, 39-41). The choice of study design depends on a number of factors such as the rarity of the disease, the number of exposures and diseases being investigated, the ability to accurately assess the exposure status, the stability of the exposure status, and the time and money available to conduct the study (Silman and Macfarlane 2002, 41-44).

2.3 Taxonomy

Taxonomy, the science of grouping multiple biological entities based on similarities and differences, is made up of classification, the orderly arrangement of taxa into larger units; nomenclature, the naming of taxa defined and delineated by the classification; and identification, the process of assigning an unknown unit to a known, and named, taxon (Cowan 1965).

Bacterial taxa, characterised using factors such as morphology, physiology, chemical and enzymatic constituents and genetics, are classified based on similarities and differences with other taxa and a precise description, or circumscription, is generated to facilitate clear differentiation from other taxa (Cowan 1965).

The Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics proposed phylogeny, or evolutionary history, as the determinant of bacterial taxonomy with the complete DNA sequence as the reference standard method (Wayne et al. 1987). DNA reassociation was considered the method that most closely approached the complete DNA

sequence, at the time, and the following species definition was proposed: “The phylogenetic definition of a species generally would include strains with approximately 70% or greater DNA-DNA relatedness and with 5°C or less ΔT_m. Both values must be considered.” (Wayne et al. 1987).

The development of nucleic acid sequencing methods has provided a powerful approach to measuring evolutionary relationships (Wilson, Carlson, and White 1977) that is superior to phenotypic information for relating and classifying bacteria because it is more readily, reliably and precisely interpreted and is innately more informative about evolutionary relationships (Woese 1987). Ribosomal RNA (rRNA) show a high level of functional constancy, occur in all organisms, are relatively large with different positions in the sequence changing at different rates and they can be sequenced directly, making them useful for evaluating phylogeny (Woese 1987).

The establishment of a rapid sequence analysis method for 16S rRNA genes and the recognition of its potential to determine the phylogenetic position of any prokaryotic organism led to the inclusion of 16S rRNA similarities in the bacterial species definition (Stackebrandt and Goebel 1994). Comparative studies had already revealed limitations associated with determining relationships at the strain level using this method and confirmed DNA reassociation was the superior method (Stackebrandt and Goebel 1994). However, since the 16S rRNA sequencing was easier to perform, it was recommended that this method be performed first and that DNA reassociation studies be performed when similarities of ≥97% were observed (Stackebrandt and Goebel 1994). Over the next two decades several reports demonstrated that phylogeny based on 16S rRNA sequence can be misleading for some Epsilonproteobacteria taxa (Hanninen et al. 2005, Hanninen et al. 2003, Harrington and On 1999, Vandamme et al. 2000).

The International Committee on Bacteriological Nomenclature recommends that the name and description of a new species conform at least to the minimal standards (if available) for the relevant taxon of bacteria before it can be published (Lapage et al. 1992). Minimal

standards published in 1994 for the family *Campylobacteraceae* (Ursing, Lior, and Owen 1994) and in 2000 for the genus *Helicobacter* (Dewhirst, Fox, and On 2000) included a polyphasic approach whereby morphological, phenotypic and genetic factors that should be included in the classification of a new species, or subspecies, within these taxa. This polyphasic approach is maintained in the new minimal standards proposed for the four genera *Arcobacter, Campylobacter, Helicobacter* and *Wolinella* which include many of the factors proposed earlier but recommending that sequence information from 16S rRNA, and additional phylogenetic markers, be used to support the position of the new taxon (On et al. 2017). The proposed new minimal standards also provide scope to include the whole genome sequence-based *in silico* analyses average nucleotide identity (ANI) and genome BLAST distance phylogeny (GBDP) in place of DNA reassociation to determine interspecific genomic relatedness (On et al. 2017).

With the aim of ensuring that phylogenetically-based taxonomic schemes also show phenotypic consistency, the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics also recommended that a genospecies[1], or genetically distinct group of strains, cannot be named as a new species until it can be differentiated from other genospecies by some phenotypic property (Wayne et al. 1987).

2.4 Epsilonproteobacteria

The bacterial class Epsilonproteobacteria[2] (Euzéby 1997, Parte 2014) contained two orders, four families and 17 genera that were validly or effectively described by December 2014 with a new genus added in 2015. The order *Campylobacterales* (Garrity, Bell, and Lilburn 2005a) is a metabolically and ecologically diverse group containing the three families *Campylobacteraceae, Helicobacteraceae* (Garrity, Bell, and Lilburn 2005c) and '*Hydrogenimonaceae'* (Parte 2014, Euzéby 1997). The family *Campylobacteraceae* consists of

[1] Also known as genomospecies

[2] http://www.bacterio.net/-classifphyla.html , accessed 20th July 2017

Gram negative, non spore-forming rods that are predominantly curved, spiral or S-shaped (Vandamme et al. 2005a). The three genera within this family are *Campylobacter, Arcobacter*, and *Sulfurospirillum* (Vandamme et al. 2005a). The family *Helicobacteraceae* consists of the morphologically, metabolically, and ecologically diverse genera *Helicobacter, Thiovulum*, *Wolinella* (Garrity, Bell, and Lilburn 2005b), *Sulfuricurvum, Sulfurimonas* and *Sulfurovum* (Parte 2014, Euzéby 1997). The family "*Hydrogenimonaceae"* (Parte 2014, Euzéby 1997) currently includes a single genus, *Hydrogenimonas* (Takai, Nealson, and Horikoshi 2004).

The order *Nautiliales* (Miroshnichenko et al. 2004) contains marine thermophiles isolated from deep sea hydrothermal vents. The family *Nautiliaceae* contained the three validly or effectively described genera *Nautilia*, *Caminibacter* and *Lebetimonas* (Nakagawa and Takai 2014) in December 2014. The genus *Cetia* was added to this family in 2015 (Grosche et al. 2015). The *Nautiliales* order also contains the three genera[3] *Nitratifractor*, *Nitratiruptor* (Nakagawa, Takai, et al. 2005) and *Thioreductor* (Nakagawa, Inagaki, et al. 2005) for which the family is uncertain.

In December 2014 there were 25 validly described species in the genus *Campylobacter* (Debruyne et al. 2010b, a, Debruyne, Gevers, and Vandamme 2008, Debruyne et al. 2009, Koziel et al. 2014, Rossi et al. 2009, Vandamme et al. 2010, Zanoni et al. 2009) and this number had increased to 30 by October 2017 (Caceres et al. 2017, Gilbert et al. 2015, Gilbert et al. 2017, Piccirillo et al. 2016, Van et al. 2016), as illustrated in Figure 2. *C. jejuni* subsp. *jejuni*, *C. coli, C. fetus* subsp. *fetus, C. hyointestinalis, C. lari* and *C. upsaliensis* have been shown to cause human gastroenteritis (Broczyk et al. 1987, Lastovica and Skirrow 2000, Salama et al. 1992, Vandamme et al. 2005b) and *C. jejuni* and *C. coli* are collectively the most commonly reported cause of human bacterial gastroenteritis in the world (Anonymous 2012), accounting for over a third of all notifications of enteric disease in New Zealand (Health Intelligence Team 2014, 2015, 2016, 2017). A number of other *Campylobacter* taxa including *C. concisus, C.*

[3] http://www.bacterio.net/-classifphyla.html, accessed 19th October 2017

curvus, C. jejuni subsp. *doylei, C. rectus, C. sputorum* and *C. ureolyticus* (On 2013) have been associated with human gastroenteritis but a causal role has yet to be established. Some species including *C. concisus, C. gracilis* and *C. ureolyticus* have been isolated (Engberg et al. 2000, Van Etterijck et al. 1996) and/or detected (Collado et al. 2013, Cornelius et al. 2012) from both healthy and diarrheic patients. *C. concisus* and *C. ureolyticus* have been shown to be genetically heterogeneous (Bullman et al. 2013, Vandamme et al. 1989) leading some to hypothesise that some strains have greater potential for causing illness than others (Aabenhus et al. 2005, Bullman et al. 2012). Although a number of methods have been proposed for differentiating pathogenic from commensal *C. concisus* (Aabenhus et al. 2005, Deshpande et al. 2013, Kaakoush, Castano-Rodriguez, et al. 2014, Kalischuk and Inglis 2011, Mahendran et al. 2013, Kaakoush et al. 2011), further research is required to establish the basis of virulence potential in this species.

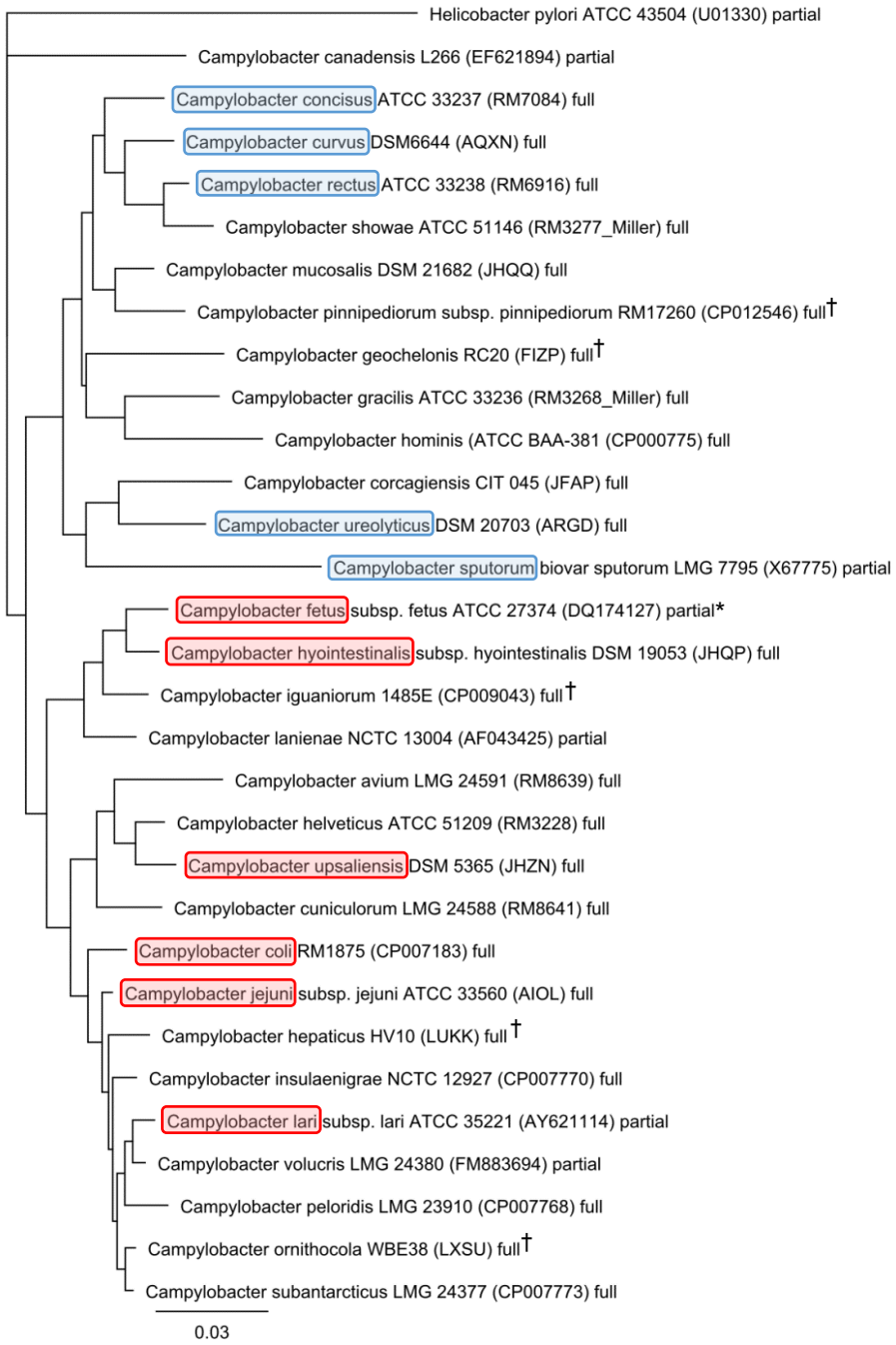


Figure 2: Phylogenetic tree for the type strains of the validly described species within the *Campylobacter* genus

Based on nucleotide sequence similarity of the 16S rRNA gene. *Helicobacter pylori* was used as the out-group. The tree was constructed in Geneious 6.1.7 (Biomatters, available at www.geneious.com) using the Tamura-Nei genetic distance model and neighbor joining with 70% similarity cost matrix, a gap open penalty of 12, a gap extension penalty of 3, and global alignment with free end gaps as the alignment type

* Type species; † validly described after December 2014; species considered to be human pathogens are highlighted in red and species that have been associated with human gastroenteritis are highlighted in blue.

The genus *Arcobacter* contained 18 validly described species by December 2014 (Lastovica, On, and Zhang 2014, Sasi Jyothsna et al. 2013), with four additional species added by October 2017 (Levican et al. 2015, Oren and Garrity 2015, Whiteduck-Leveillee et al. 2015, Zhang et al. 2016), as illustrated in Figure 3. *A. butzleri* and *A. cryaerophilus* are considered human pathogens by the International Committee for the Microbiological Safety of Foods (Intenational Commission on Microbiological specifications for Foods 2002); and both *A. skirrowii* and *A. thereius* have been associated with human gastroenteritis (Van den Abeele et al. 2014, Vandamme et al. 2005c, Wybo et al. 2004).

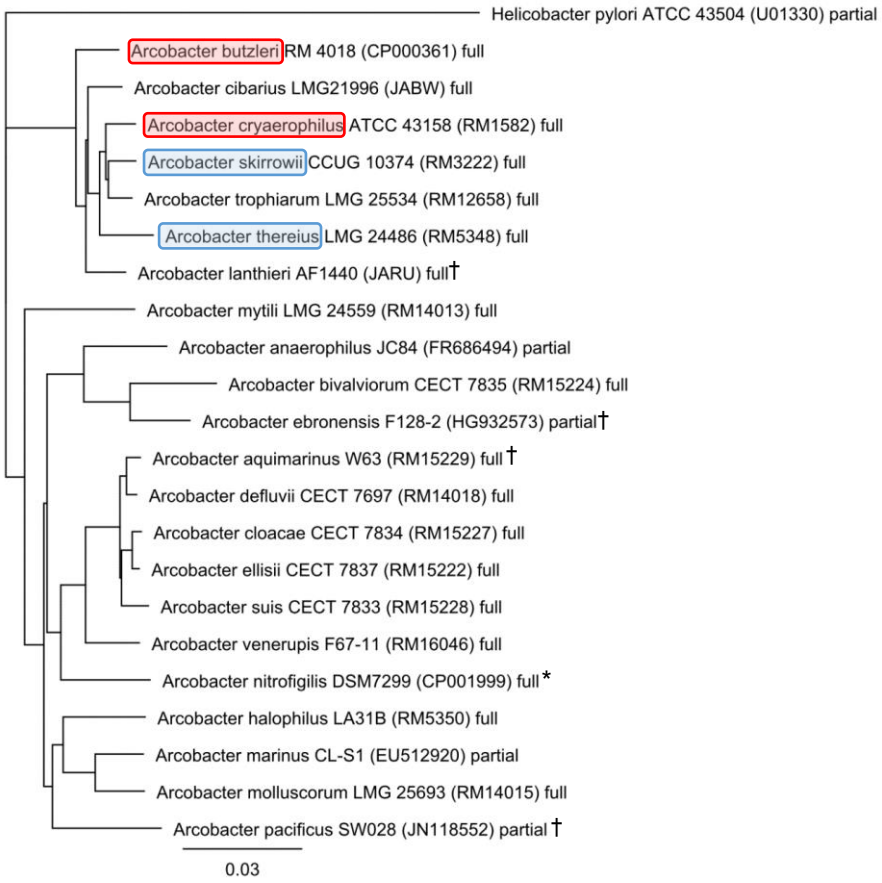


Figure 3: Phylogenetic tree for the type strains of the validly described species within the *Arcobacter* genus

Based on nucleotide sequence similarity of the 16S rRNA gene. *Helicobacter pylori* was used as the out-group. The tree was constructed in Geneious 6.1.7 (Biomatters, available at www.geneious.com) using the Tamura-Nei genetic distance model and neighbor joining with 70% similarity cost matrix, a gap open penalty of 12, a gap extension penalty of 3, and global alignment with free end gaps as the alignment type

* Type species; † validly described after December 2014; species considered to be human pathogens are highlighted in red and species that have been associated with human gastroenteritis are highlighted in blue.

In December 2014 there were 25 validly described species in the genus *Helicobacter*, with an additional 7 species added by October 2017 (Oren and Garrity 2017, Parte 2014, Shen, Feng, et al. 2016, Shen et al. 2015, Shen, Mannion, et al. 2016)[4], as illustrated in Figure 4. *H. pylori* causes gastritis, peptic ulcers and gastric malignancy and is a class 1 carcinogen

[4] http://www.bacterio.net/helicobacter.html, accessed 19th October 2017

(Anonymous 1994). Seven additional species (*H. bilis, H. bizzozeronii, H. canadensis, H. canis, H. cinaedi, H. fennelliae,* and *H. pullorum*) are suspected human pathogens (Melito et al. 2001, On et al. 2005). *H. bilis* has been isolated from bile and gall bladder samples; *H. bizzozeronii* has been isolated from a gastric biopsy; and *H. canadensis, H. canis, H. cinaedi, H. fennelliae,* and *H. pullorum* have been associated with diarrhoea (On et al. 2005).

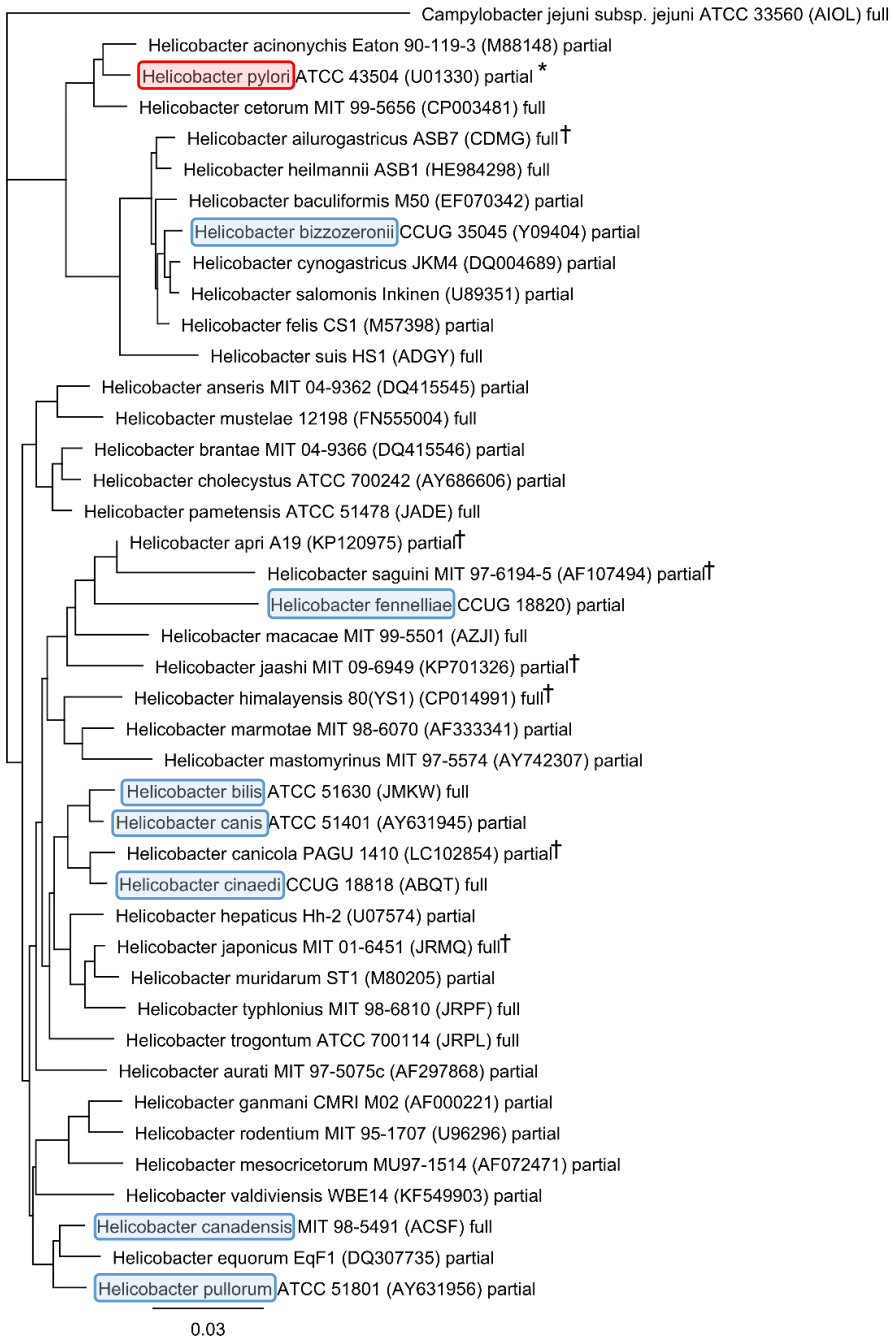


Figure 4: Phylogenetic tree for the type strains of the validly described species within the *Helicobacter* genus

Based on nucleotide sequence similarity of the 16S rRNA gene. *Campylobacter jejuni* was used as the out-group. The tree was constructed in Geneious 6.1.7 (Biomatters, available at www.geneious.com) using the Tamura-Nei genetic distance model and neighbour joining with 70% similarity cost matrix, a gap open penalty of 12, a gap extension penalty of 3, and global alignment with free end gaps as the alignment type

* Type species; † validly described after December 2014; species considered to be human pathogens are highlighted in red and species that have been associated with human gastroenteritis are highlighted in blue.

The remaining Epsilonproteobacterial taxa, including the remaining *Helicobacteraceae* and *Campylobacteraceae* genera and all of the genera within the *Nautiliales* order, have only been isolated from environmental or animal sources, are associated with non-gastrointestinal illness, or have not yet been associated with human illness (Alain et al. 2009, Alain et al. 2002, Inagaki et al. 2003, Inagaki et al. 2004, Kodama, Ha, and Watanabe 2007, Kodama and Watanabe 2004, Kroger et al. 2005, Labrenz et al. 2013, Luijten et al. 2003, Miroshnichenko et al. 2002, Miroshnichenko et al. 2004, Nakagawa, Inagaki, et al. 2005, Nakagawa, Takai, et al. 2005, Perez-Rodriguez et al. 2010, Robertson et al. 2005, Smith et al. 2008, Sorokin, Tourova, and Muyzer 2013, Stolz et al. 2005, Takai et al. 2005, Takai, Nealson, and Horikoshi 2004, Takai et al. 2006, Voordeckers, Starovoytov, and Vetriani 2005).

A recent publication proposes reclassification of the class Epsilonproteobacteria to the phylum Epsilonbacteraeota with the current class name being replaced with Campylobacteria to remove the reference to Proteobacteria (Waite et al. 2017). This proposal would also introduce a new class called Desulfurellia incorporating the order Desulfurellales from the current Deltaproteobacteria class; introduce the new families *Arcobacteraceae, Nitratriuptoraceae, Sulfurospirillaceae, Sulfurovaceae* and *Thiovulaceae* within Campylobacterales; new family *Thioreductoraceae* within Nautiliales; and the new family *Hippeaceae* within Desulfurellales (Waite et al. 2017). Some genera would move to these new families as illustrated in Figure 5.

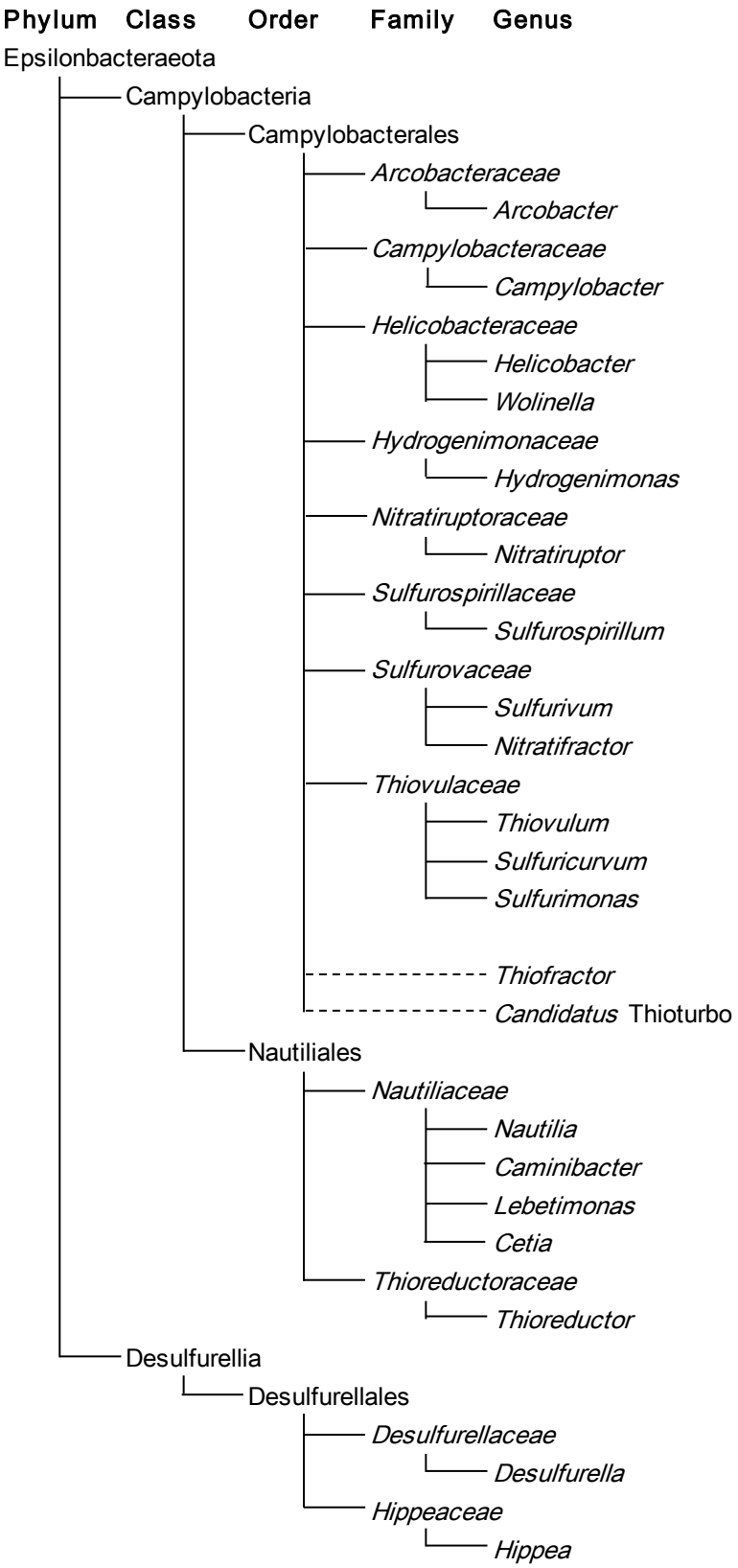


Figure 5: Organisation of the new Phylum Epsilonbacteraeota as proposed by Waite et al. 2017 Front Microbiol 8:682

2.5 Culture Methods

2.5.1 Isolation Methods

Culture from stools is the primary diagnostic tool for suspected bacterial gastroenteritis (Zhang, Morrison, and Tang 2015). This normally involves the use of selective agars followed by morphologic, biochemical and serologic testing in order to identify and confirm the suspected isolate which is time-consuming, labour-intensive, expensive and requires 3 – 5 days (Dunbar, Zhang, and Tang 2013).

Clinical laboratories routinely test gastroenteritis-associated stool samples for *Campylobacter* species but the method employed by most laboratories uses selective media and growth conditions that limit the species isolated to *C. jejuni* subsp. *jejuni* (hereafter called *C. jejuni* unless otherwise stated) and *C. coli* (Lindblom et al. 1995, Parker et al. 2007). Isolation is not routinely attempted for any non-*Campylobacter* members of the Epsilonproteobacterial class because of the varied and complex nutrient and atmospheric requirements, coupled with often slow growth rates, of other gastroenteritis-associated Epsilonproteobacteria (On et al. 2005, Vandamme et al. 2005b, c) make isolation of species within the class challenging. Passive filtration onto antibiotic-free blood agar and incubation in a microaerophilic atmosphere enriched with hydrogen has been shown to facilitate the isolation of a broader range of Epsilonproteobacterial species (Lastovica and Le Roux 2000). Maximum recovery, of at least some species, is achieved by including both selective and non-selective filtration methods (Van Etterijck et al. 1996).

When these enhanced culture methods were employed, several non-*C. jejuni/C. coli* Epsilonproteobacterial species, including *C. concisus, C. upsaliensis, A. butzleri* and *C. jejuni* subsp. *doylei,* were isolated in significant numbers from gastroenteritis cases in several countries (Aabenhus et al. 2005, Engberg et al. 2000, Goossens et al. 1990, Lastovica 2006, Lastovica and Le Roux 2000, Lindblom et al. 1995, Musmanno et al. 1998, Nielsen, Engberg, et al. 2013b, Taylor et al. 1991, Vandenberg et al. 2004). The only New Zealand study using the

passive filtration method isolated *C. concisus* from only 1 of 200 diarrheic samples (Keenan et al. 2014) although it is likely that the atmosphere in this study was not hydrogen-enriched.

2.5.2 Identification Methods

The process of identification involves matching data sets derived from an unknown isolate to those of defined taxa (On 2005). An isolate is considered identified only if an acceptable level of matching is achieved (On 2005). For *Arcobacter, Campylobacter* and *Helicobacter* species, this process can be problematic due to their relatively inert nature and complex taxonomy (On 2013). Traditional phenotypic methods involving mostly biochemical tests have been a common approach to identify *Arcobacter*, *Campylobacter* and *Helicobacter* species (Collado and Figueras 2011, On and Holmes 1995, Owen 1998). However, variations in laboratory protocols (On 1996) and reliance on a few phenotypic tests can lead to misidentifications (On 2005). This is further complicated by phenotypic variation within a species as exemplified by *C. jejuni* strains that lack the ability to hydrolyse hippurate, urease-negative *H. pylori* and urease-positive *C. sputorum* (On 2001). A well-documented example of misidentification within Epsilonproteobacteria involved the reported isolation of *C. mucosalis* from stool samples associated with childhood enteritis (Figura et al. 1993) which were subsequently shown to be *C. concisus* (Lastovica et al. 1993, On 1994, Lastovica et al. 1994). Alternative phenotypic methods such as cellular fatty acid profiling, whole-cell protein profiling and mass spectrometry (including matrix assisted laser desorption ionisation time-of-flight [MALDI-TOF]) have also been applied to some species in each of these three genera (On 2005, Baele et al. 2008, Brondz and Olsen 1991, De Smet et al. 2011, Debruyne et al. 2010b, Donachie et al. 2005, Fowsantear et al. 2014, Geis et al. 1990, Houf et al. 2009, Kaur et al. 2011, Kim, Hwang, and Cho 2010, Suerbaum et al. 1992, Vandamme et al. 2010, Alispahic et al. 2010, Bessede et al. 2011, Levican et al. 2012, Levican, Collado, and Figueras 2013, Martiny et al. 2011, Murray 2010, Taniguchi et al. 2014, Winkler, Uher, and Cepa 1999).

Amplification and detection of taxon-specific DNA segments using the polymerase chain reaction (PCR) has emerged as a popular method for molecular identification of *Arcobacter* (Collado and Figueras 2011), *Campylobacter* (On 2013) and *Helicobacter* (On et al. 2005) species. These PCR methods, which produce a presence/absence type result, are relatively quick and easy to perform.

2.6 Antigen Detection Methods

Antigen detection methods use antibodies attached to a solid surface to capture cells by binding to structures on the cell surface. A second antibody, with an attached reporter system, is then used to visualise the interaction. Assays involving an enzyme as the reporter system are generally undertaken in microtitre plates by trained personnel and are called enzyme immunoassays (EIA), enzyme-linked immunosorbent assays (ELISA) or microplates and generally require a spectrophotometer (plate reader) for detection (Lequin 2005). Immunochromatographic assays (ICA), or lateral flow assays, which often use coloured particles as the reporting system, are very simple and easy to use tests that are rapid and do not require trained personnel or specialised equipment (Jani et al. 2002).

A range of antigen detection methods, including both EIA and ICA formats are commercially available for the detection of *C. jejuni/C. coli* (Fitzgerald et al. 2016, Gomez-Camarasa et al. 2014) and *H. pylori* (Canadian Agency for Drugs and Technologies in Health 2015) in human stools. An alternative assay, which uses aptamers rather than antibodies to capture the cell surface structures and quantum dots for the report system, has been developed for the detection of *C. jejuni* in foods (Bruno et al. 2009). Aptamers are synthetically generated nucleic acids with antibody-like binding ability that are developed through iterative cycles of affinity selection and PCR amplification (Bruno et al. 2009).

The assays targeting *C. jejuni/C. coli* generally have high specificities (>95%) and negative predictive values (NPV) (>99%), but the sensitivities (<90%) and positive predicative values (<80%) are generally lower (Fitzgerald et al. 2016, Gomez-Camarasa et al. 2014). The

results for the *H. pylori* stool antigen tests are more variable (Canadian Agency for Drugs and Technologies in Health 2015). These methods are rapid, and generally quite specific, but *C. jejuni/C. coli* and *H. pylori* are the only Epsilonproteobacterial species for which methods of this kind have been published or commercialised.

2.7 Genetic Methods for Detection

Genetic methods offer the potential to concurrently detect a broad range of bacterial species in a manner that is not biased by culture conditions. A variety of methods have been published for the genetic detection of the established or potential human pathogenic Epsilonproteobacterial species in human stool samples. These methods include loop-mediated isothermal amplification (LAMP, (Minami et al. 2006, Miyagawa et al. 2008, Yamazaki et al. 2008)), PCR (Ceelen et al. 2005, Collado et al. 2013, Oyama et al. 2012, Rimbara, Sasatsu, and Graham 2013, Yamazaki-Matsune et al. 2007) and PCR-denaturing gradient gel electrophoresis (PCR-DGGE, (Cornelius et al. 2012, Vandenberg et al. 2013)) assays. These genetic methods generally detect either a single taxon, or less than 10 taxa per reaction. The PCR-DGGE method, although useful for identifying multiple Epsilonproteobacterial taxa in a single assay, is time-consuming and not suitable for routine use. DNA microarray assays have been reported for the detection of a range of gastrointestinal pathogens (Donatin et al. 2013, You et al. 2008) but they have been designed to detect only a small number of Epsilonproteobacterial species (Donatin et al. 2013).

Multiplex ligation-dependent probe amplification (MLPA) is a modification of PCR (Figure 6) that allows up to 40 genes to be targeted within a single reaction (Schouten et al. 2002). Each MLPA probe consists of two oligonucleotides that are ligated to each other when hybridised to a target sequence (Schouten et al. 2002). All MLPA probes have the same 5' and 3' ends permitting simultaneous PCR amplification with a single primer pair (Schouten et al. 2002). The MLPA probes are generally designed to have unique amplification product lengths allowing recognition of each target sequence on the basis of the size of the amplification

products (Schouten et al. 2002). MLPA assays can be designed to be undertaken in a single working day and use relatively basic molecular biology equipment (Cornelius et al. 2014). A MLPA assay has been reported for the simultaneous detection of 13 pathogenic bacteria, including *C. jejuni* from foods where the products are similar lengths and are separated using single capillary electrophoresis-single strand conformation polymorphism (CE-SSCP) (Kim et al. 2016).

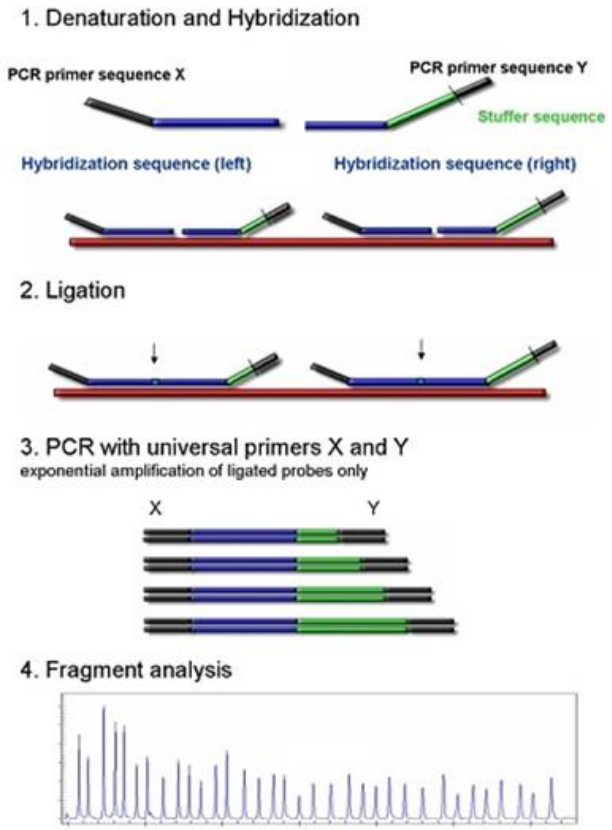


Figure 6: Illustration of Multiplex Ligation-dependent Probe Amplification (MLPA)
Courtesy of MRC-Holland (http://mlpa.com)

2.8 Whole Genome Sequencing

The whole genome can be said to be the ultimate taxonomic reference marker (On 2005). Dideoxy chain-termination, or Sanger, sequencing (Sanger, Nicklen, and Coulson 1977) was the dominant method of generating DNA sequence information for over 25 years (Liu et al. 2012). This method can output 2.88 Mb per day with a read length of 900 bases (Liu et al. 2012) but is too laborious and expensive for routine sequencing of anything more than a few genes

(Moorthie, Mattocks, and Wright 2011). During the Sanger era a total of 192 bacterial whole genome sequences (WGS) had been fully completed and published (Forde and O'Toole 2013).

The 2001 human genome project stimulated the development of powerful novel sequencing equipment to increase speed and accuracy, while simultaneously reducing cost and manpower (Collins, Morgan, and Patrinos 2003, Liu et al. 2012). The resulting next generation sequencers (NGS) are represented by Roche's 454, Applied Biosystems' Sequencing by Oligo Ligation Detection (SOLiD), Ion Torrent's Ion Personal Genome Machine (IPGM) and Illumina's HiSeq, NextSeq, MiSeq and MiniSeq (Goodwin, McPherson, and McCombie 2016, Liu et al. 2012, Mardis 2013). These machines sequence millions of small fragments of DNA in parallel (Behjati and Tarpey 2013) thus providing higher throughput at a cost lower than Sanger sequencing (Liu et al. 2012). The accuracy of the Illumina platform is now approaching that of Sanger sequencing (Lynch et al. 2016). These next generation, or massively parallel, sequencers produce sequence information using a variety of techniques including pyrosequencing (454), ligation and two-base calling (SOLiD), pH changing monitoring (IPGM) and sequencing by synthesis (Illumina) and have been reviewed elsewhere (Liu et al. 2012, Mardis 2013). The 454 and SOLiD machines are no-longer being produced.

In contrast to the short ≤300 bp reads provided by the IPGM and Illumina sequencers, longer reads of between 1,000 and 100,000 bp are generated via the single-molecule sequencers from Pacific Biosciences (PacBio RS II and PacBio Sequel) and Oxford Nanopore Technologies (MinION and PromethION) (Goodwin, McPherson, and McCombie 2016, Taboada et al. 2017). The strength of these long read sequencers, which have error rates of greater than 10% (Goodwin, McPherson, and McCombie 2016, Taboada et al. 2017), lies in their contribution to generating "scaffolds" used for inter-connecting high quality contiguous sequences (contigs) generated by short read technologies; in combination they permit efficient reconstruction of genomes (Taboada et al. 2017).

Mapping of NGS data to a reference genome of a closely related strain and performing variant calling can provide valuable information about sequence variations between closely

related strains (Lynch et al. 2016). However, for most bacterial genome projects, the first step is to reconstruct a genome, or at least generate large contigs (Forouzan et al. 2017), from raw sequencing data without the aid of a reference, a process called *de novo* assembly (Lynch et al. 2016). Over 20 assembly software programs or pipelines have been applied to the *de novo* assembly of bacterial genomes (Dark 2013, Forouzan et al. 2017, Lynch et al. 2016, Magoc et al. 2013, Salzberg et al. 2012, Scott and Ely 2015). Most of these assemblers are freely available (open source), can incorporate raw sequence data from multiple sequencing technologies and use either overlap-layout-consensus (OLC) or de Bruijn graph algorithms (Dark 2013, Lynch et al. 2016). The OLC strategy involves organising reads into graph structures with each read being a node which is connected by an edge to other overlapping reads, and de Bruijn graph assemblers first partition the reads into overlapping subsequences of length k, called kmers, which are used to create the nodes (Lynch et al. 2016).

Genome annotation is the preliminary link between nucleotide sequence and biological function (Aken et al. 2016) and involves identifying biologically important features and attaching descriptive information to these features (Lynch et al. 2016). Gene prediction is the first step in genome annotation and involves the locating of protein coding genes (CDS) and noncoding genes (including tRNA and rRNA) (Lynch et al. 2016). Programs designed to predict bacterial genes are either intrinsic or extrinsic. Intrinsic gene finders, such as GeneMark.hmm (Lukashin and Borodovsky 1998), GLIMMER (Delcher et al. 1999) and Prodigal (Hyatt et al. 2010), attempt to identify coding sequences based solely on the information contained within the newly sequenced and assembled genome or contig (Lynch et al. 2016). Extrinsic gene finders, such as ORPHEUS (Frishman et al. 1998) and CRITICA (Badger and Olsen 1999), use a database of previously identified and verified bacterial protein coded sequences to aid in the identification of genes in a new genome (Lynch et al. 2016).

The CDSs in a bacterial genome defines its biology so it is of great interest and importance to characterise the functions of these CDSs, especially when predicting or assessing pathogen virulence or risk (Lynch et al. 2016). Characterisation of the biological

function of CDSs requires *in vitro* analysis which is time-consuming and often challenging to perform, so *in silico* methods of functional annotation are used as a first step in biological function prediction. Modern functional annotation approaches use a combination of sequence similarity searches (using the BLAST family of similarity searching programs), Hidden Markov model (HMM)-based searches, and a variety of biochemical property searches, such as Pfam[5] and PROSITE[6], to infer the function of genes in newly sequenced genomes (Lynch et al. 2016).

Traditionally, genome annotation relied heavily on manual expert curation, but this is not feasible with the exponential growth of sequence data (Chen et al. 2016) and concomitant deposition of draft prokaryotic genomes in databases (see below). Rapid and accurate automatic genome annotation systems have been developed that compile a multitude of tools into coordinated pipelines that remove much of the complexity of performing bacterial genome annotation (Lynch et al. 2016). Web-based pipelines (Lynch et al. 2016, Vallenet et al. 2017) include Rapid Annotations using Subsystems Technology[7] (RAST, (Aziz et al. 2008)), Integrated Microbial Genomes Expert Review[8] (IMG-ER, (Markowitz et al. 2009)), Pathosystems Resource Integration Center[9] (PATRIC (Wattam et al. 2014, Wattam et al. 2017), The SEED[10] (Overbeek et al. 2005), Ensembl Genomes[11] (Kersey et al. 2016), Evolutionary Genealogy of Genes: Nonsupervised Orthologous Groups (eggNOG[12] (Huerta-Cepas et al. 2016, Jensen et al. 2008)), OrthoDB[13] (Kriventseva et al. 2015) and Magnifying Genomes (MaGe)/MicroScope[14] (Vallenet et al. 2009, Vallenet et al. 2006). Downloadable pipelines,

[5] https://pfam.xfam.org

[6] https://prosite.expasy.org

[7] http://rast.nmpdr.org/

[8] https://img.jgi.doe.gov/er

[9] https://www.patricbrc.org

[10] http://theseed.org/wiki/Main_Page

[11] http://ensemblgenomes.org/

[12] http://eggnogdb.embl.de/#/app/home

[13] https://www.orthodb.org/v8/

[14] http://www.genoscope.cns.fr/agc/microscope

including Prokka (Seemann 2014) and Do-It-Yourself Annotator (DIYA (Stewart, Osborne, and Read 2009)), can take advantage of both high performance local workstations and also at times where data privacy issues are critical (Lynch et al. 2016, Vallenet et al. 2017). In addition, some genome annotation tools such as BLAST2GO (Gotz et al. 2008) are available as both online and standalone versions and the National Center for Biotechnology Information (NCBI) now offers automated annotation, as part of the genome submission process, using its Prokaryotic Genome Annotation Pipeline (PGAP, (Tatusova et al. 2016)).

The contribution of NGS to bacterial genomes is demonstrated by the additional 1566 complete bacterial WGS which were published and deposited in online databases during the seven years between the introduction of this technology in 2005 and October 2012 (Forde and O'Toole 2013). These online databases include the NCBI's GenBank[15], the DNA DataBank of Japan (DDBJ[16]) and the European Nucleotide Archive[17] within the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI[18]) which together make up the International Nucleotide Sequence Database Collaboration (INSDC[19]). These databases provide a repository of genomic sequences, in draft and complete form, as well as the sequence read archive (SRA[20]) where raw data from a range of sequencing platforms can be stored and shared. In addition, there are other publicly available online data repositories including PubMLST[21], EnteroBase[22] and DRYAD[23].

[15] www.ncbi.nlm.nih.gov/genbank

[16] http://www.ddbj.nig.ac.jp/

[17] http://www.ebi.ac.uk/ena

[18] http://www.ebi.ac.uk/

[19] www.insdc.org

[20] for example the SRA at NCBI https://www.ncbi.nlm.nih.gov/sra

[21] https://pubmlst.org/

[22] https://enterobase.warwick.ac.uk/

[23] https://datadryad.org

2.9 Core and Pan-genomic Analysis

As the cost of NGS decreased it became feasible to generate WGS for multiple strains within a species. The observation that the gene repertoire of a bacterial species was much larger than the genes carried by a single strain led to the concept that a bacterial species could be represented by its pan-genome, which includes a core genome and a dispensable or accessory genome (Tettelin et al. 2005). The core genome consists of genes shared by all strains; and the accessory, or dispensable, genome includes genes that are present in some, but not all, strains (Tettelin et al. 2005). Figure 7 is an illustration of a hypothetical taxon of three strains containing an equal number of genes where the core genome is 500 genes, the accessory genome contains 2050 genes and the pan-genome contains 2550 genes.

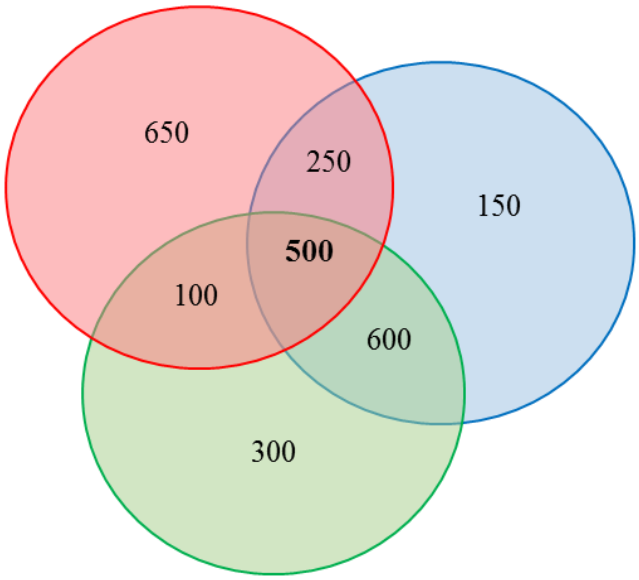


Figure 7: Illustration of a Venn Diagram for a Set of Three Isolates each with 1500 Genes The centre intersection (500) represents the core genome of this set of three isolates. The isolate specific (650, 150 and 300) and two-isolate intersections (250, 600 and 100) combined (2050) represent the accessory genome of the set. All genes from both the core and accessory genomes (2550) is the pan-genome.

Pan-genomic analysis can take many forms including studies evaluating the conservation of gene order (Lee, Grasso, and Sharlow 2002), identifying the genes that make up the core genome and accessory genomes (Laing et al. 2010), and estimating the size of a species' core genomes and pan-genomes (Tettelin et al. 2005). A growing number of free software programs

have been published for identifying core and accessory genes from multiple bacterial WGS. Web-based pan-genome analysis programs include Efficient Database framework for comparative Genome Analyses using BLAST score Ratios (EDGAR), Prokaryotic-genome Analysis Tool (PGAT), panX and mGenomeSubtractor (Blom et al. 2009, Brittnacher et al. 2011, Ding, Baumdicker, and Neher 2018, Shao et al. 2010). The utility of EDGAR and PGAT is, however, limited to the genomes in their databases. Programs available only as standalone pipelines include Gegegees, Integrated Toolkit for Exploration of microbial Pan-genomes (ITEP), Bacterial Pan Genome Analysis (BPGA), GET_HOMOLOGUES, Accessory Genome Constellation Network (AcCNET), Roary and the associated Scoary, PanCoreGen, Large Scale BLAST Score Ratio (LS-BSR) and Pan-Genomics Tool (PanGeT) (Agren et al. 2012, Benedict et al. 2014, Brynildsrud et al. 2016, Chaudhari, Gupta, and Dutta 2016, Contreras-Moreira and Vinuesa 2013, Lanza et al. 2017, Page et al. 2015, Paul et al. 2015, Sahl et al. 2014, Yuvaraj et al. 2017). There are also several programs that are available as both web-based and standalone tools such as Panseq, PAN-genome analysis based on FUNctional PROfiles (PanFunPro), Spine and AGEnt, and Pan-Genomes Analysis Pipeline (PGAP) and PanWeb (Laing et al. 2010, Lukjancenko et al. 2013, Ozer, Allen, and Hauser 2014, Pantoja et al. 2017, Zhao et al. 2012).

The success of any pan-genomic analysis is dependent on the quality of the input data. High quality assemblies can only be obtained if the input sequence reads are of suitable quality. This can be assured by evaluating the base call quality scores, read signal intensity, depth and uniform of coverage and the balance between multiplexed libraries (Lynch et al. 2016). In addition, cleaning of the reads may be necessary to remove, mask or trim low-quality reads and to remove adapters and other sequencing artefacts (Lynch et al. 2016). It is also advisable to detect and remove other contaminating DNA. Sources of these unwanted sequences includes control DNA, such as the bacteriophage PhiX used by Illumina, which is used as a quality and calibration control and to provide sufficient sample complexity to ensure optimal base calling

(Mukherjee et al. 2015). Mixed or contaminated starting DNA is also a source of unwanted sequences (Lynch et al. 2016).

Genome assembly is a complex computational challenge and no algorithm is guaranteed to accurately and completely reconstruct genomes from the short sequences generated by most modern sequencers (Simpson and Pop 2015). Determining whether an assembly is correct, or comparing the quality of different assemblies of the same data, is difficult given that the correct answer is usually not known (Nagarajan and Pop 2013). Finishing, or completing, a genome is the process of closing all the contig gaps, correcting any introduced errors and confirming low coverage regions through PCR and cloning experiments (Ricker, Qian, and Fulthorpe 2012). This process takes months to years so this is rarely carried out due to the effort required and because many sequencing projects are evaluating small differences between closely related genomes (Ricker, Qian, and Fulthorpe 2012). Programs such as QUAST[24] (Gurevich et al. 2013) provide some measures of assembly quality but they cannot detect small local errors such as single nucleotide errors or rearrangements and do not provide a measure of acceptable quality where a single assembly is available. Recent evaluations of assemblers have established that the vast majority of protein-coding genes will be contained wholly within contigs generated from single library assemblies (Magoc et al. 2013) and these assemblies are accurate enough for gene level analysis (Forouzan et al. 2017). Complete genomes provide the best starting information for pan-genomic analysis but as long as finishing genomes is out of the reach of most laboratories, draft genomes provide an adequate starting point for these analyses. Draft genome-based pan-genomic analyses that evaluate the presence or absence of genes within taxa require additional analyses to confirm bacterial strains do not carry genes that are absent from the associated draft genomes.

The success of the entire genomic enterprise depends on reliable genome annotation (Galperin et al. 2017). Errors in genome annotation can arise when random open reading

[24] http://bioinf.spbau.ru/quast

frames, predicted from start and stop codons, are misidentified as genes; sequence similarity is mistaken for functional similarity; and where reference databases contain incorrect annotations (Stothard and Wishart 2006). Examples of such reference databases include Clusters of Orthologous Groups (COG[25], (Tatusov, Koonin, and Lipman 1997, Tatusov et al. 2001)), GenBank[26] (Benson et al. 2017), InterPro[27] (Finn et al. 2017), Pfam[28] (Finn et al. 2016), RefSeq[29] (O'Leary et al. 2016, Klimke et al. 2009) and UniProt[30] (The UniProt Consortium 2017). Annotation challenges are continually being addressed through the development of more completely annotated databases, better gene prediction algorithms and more sensitive sequence comparison methods so the most useful annotations are usually derived from the most recently developed (or updated) annotation pipelines (Stothard and Wishart 2006).

The core genome of a bacterial species is responsible for the basic biology and major phenotypic traits of the species (Chaudhari, Gupta, and Dutta 2016) and sequence variation within these core genes has been used to evaluate the phylogenetic relationship of strains within bacterial species (Carter 2017, Cody et al. 2017). Conversely, the identification of novel accessory genes has application in characterising novel metabolic pathways, virulence attributes, adaptation to environmental life, host associations and providing molecular fingerprinting targets useful in epidemiological and population genetics studies (Baig et al. 2015, Benedict et al. 2014, Chen and Shapiro 2015, Laing et al. 2010, Salipante et al. 2015, Thepault et al. 2017).

[25] http://www.ncbi.nlm.nih.gov/COG

[26] www.ncbi.nlm.nih.gov/genbank

[27] http://www.ebi.ac.uk/interpro/

[28] http://pfam.janelia.org/

[29] http://www.ncbi.nlm.nih.gov/refseq/

[30] https://www.uniprot.org/

2.10 Conclusions

The Epsilonproteobacteria class contains several known human pathogens but suboptimal identification and isolation methods result in many pathogens being underrepresented. To help provide evidence towards inferring causation, methods unbiased by current culture conditions need to be developed. The number of WGS now available for taxa within the Epsilonproteobacterial class provide the starting material for pan-genomic analysis of the three genera that contain taxa recognised as, or suspected of being, human enteric pathogens. These pan-genomic analyses have the potential to identify taxon-specific genes that can be used to develop a rapid and cost-effective method that simultaneously detects a range of Epsilonproteobacterial taxa. Application of the pan-genome guided genetic method to samples from an epidemiological case control study will facilitate a greater understanding of the role non-*C. jejuni/coli* Epsilonproteobacteria play in human gastroenteritis.

Chapter 3: Comparative Analysis of *Campylobacter concisus*

3.1 Abstract

Campylobacter concisus is a phenotypically and genetically heterogeneous species that has been isolated or detected in a range of animal species and a variety of sites in the human body, including stool samples from healthy and diarrhoeic individuals. Draft genomes for eight well-characterised *C. concisus* strains were generated to supplement the information provided by the 10 publicly available genomes for this species and four genomes shared by Dr Mohsina Huq. Nine genomes representing related species were included to provide context. Comparisons were made using Average Nucleotide Identity (ANI), Tetranucleotide frequency (Tetra), Genome BLAST Distance Phylogeny (GBDP), Feature Frequency Profiling (FFP), ribosomal Multi-Locus Sequence Typing (rMLST), OrthoMCL, Clusters of Orthologous Genes (COGs) analysis and 16S rRNA gene sequence. All eight analyses clustered the *C. concisus* genomes into the same two groups of genomes represented by the type strain ATCC 33237^T (GS1) and CCUG 19995 (GS2). The two *C. concisus* genomospecies were well separated from the nine genomes representing related species. GS1 and GS2 exhibited differences in GC content with medians of 37.585 and 39.455, respectively. The genomospecies are consistent with published genomospecies where strains were shared and are supported by DNA reassociation results. Pan-genomic analysis identified genes specific to GS1 and GS2. Whole genome sequence (WGS) data and genomic species identification methods have provided additional support for the existence of genomospecies within *C. concisus.* These data support genomic species identification methods as a viable option for tentatively assigning partially characterised isolates to a taxonomic group.

3.2 Introduction

Classification is the division of taxonomy concerned with arranging taxa into larger units (Cowan 1965). The circumscription, or the precise definition of how each subdivision is separated, is necessary so that the units can be clearly recognised (Cowan 1965). The Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics proposed that phylogeny was the determinant of bacterial taxonomy and that complete DNA sequence would be the reference standard for establishing phylogeny (Wayne et al. 1987). The authors noted that the species is the only taxonomic unit that could be phylogenetically defined and that DNA reassociation most closely represented the complete DNA sequence (Wayne et al. 1987). The following definition of a species was recommended: “The phylogenetic definition of a species generally would include strains with approximately 70% or greater DNA-DNA relatedness and with 5°C or less ΔT_m. Both values must be considered.” (Wayne et al. 1987).

The development of methods allowing sequencing of nucleic acids has provided a powerful approach to measuring evolutionary relationships (Wilson, Carlson, and White 1977) that is superior to phenotypic information for relating and classifying bacteria because it is more readily, reliably and precisely interpreted and is innately more informative about evolutionary relationships (Woese 1987). Ribosomal RNA (rRNA) show a high level of functional constancy, occur in all organisms, are relatively large with different positions in the sequence changing at different rates and they can be sequenced directly, making them useful for evaluating phylogeny (Woese 1987).

The establishment of a rapid sequence analysis of the 16S rRNA genes and the recognition of its potential to determine the phylogenetic position of any prokaryotic organism led to an evaluation of the 16S rRNA similarities in the standing bacterial species definition (Stackebrandt and Goebel 1994). Comparative studies have revealed limitations of the sequence analysis of this conserved gene in determining relationships at the strain level and confirmed DNA reassociation as the superior method (Stackebrandt and Goebel 1994). Since

the 16S rRNA nucleotide sequence is easier to determine than hybridisation between DNA strands, it was recommended that 16S rRNA similarities of at least 97% be used to recognise which DNA reassociation studies need to be performed (Stackebrandt and Goebel 1994). More recently, a 16S rRNA-based operational definition of species, with a diversity threshold of 1 – 1.3% has been proposed (Pei et al. 2010). In a study of intragenomic 16S rRNA diversity, Tian et al. (2015) found that 446 (20.8%[31]) of 2143 finished prokaryotic genomes available from NCBI on 22 June 2014 had a single copy of the 16S rRNA gene. In addition, of the 1697 genomes that contained multiple copies of this gene, 633 (37.3%) had identical sequences, 925 (54.5%) had lowest intragenomic 16S rRNA gene similarity of between 99% and 100%, 57 were between 98.7% and 99% (the borderline diversity for species definition), and 80 were <98.7% (Tian et al. 2015). These data suggest that intragenomic variation in 16S rRNA sequence would only rarely result in misclassification.

The 16S rRNA has been found to poorly discriminate between several species including *Yersinia pseudotuberculosis* and *Y. pestis* (Ibrahim et al. 1993, Kotetishvili et al. 2005), *Escherichia coli* and *Shigella* species (Lukjancenko, Wassenaar, and Ussery 2010), members of the *Bacillus cereus* complex (Jimenez et al. 2013), as well as *C. jejuni*, *C. coli* (Hansson et al. 2008) and nontypical *C. lari* (Gorkiewicz et al. 2003, Korczak et al. 2006). In contrast, significant variation in 16S rRNA gene sequence have been observed between members of several *Campylobacter* species and subspecies (Harrington and On 1999, Korczak et al. 2006). Discordance between 16S rRNA gene sequence and other genetic or phenotypic methods of classifying species has also been observed within *Helicobacter* (Dewhirst et al. 2005, Hanninen et al. 2003, Vandamme et al. 2000).

With the aim of ensuring that phylogenetically-based taxonomic schemes also show phenotypic consistency, the Ad Hoc Committee on Reconciliation of Approaches to Bacterial

[31] Some of the percentages from this study have been corrected after consulting the corresponding author

Systematics also recommended that “a distinct genospecies[32] that cannot be differentiated from another genospecies on the basis of any known phenotypic property not be named until they can be differentiated by some phenotypic property” (Wayne et al. 1987).

The International Committee on Bacteriological Nomenclature was set up in 1930 and a Code of Nomenclature (The Code) was approved in 1947, revised, annotated and published in 1958 (Cowan 1965), revised again in 1976 (Murray and Schleifer 1994) and, most recently, 1990 (Lapage et al. 1992). The majority of The Code relates to the names given to a new taxon, but Recommendation 30b states that “before publication of the name and description of a new species, the examination and description should conform at least to the minimal standards (if available) required for the relevant taxon of bacteria” (Lapage et al. 1992). Recommendations for bacterial classification and identification are generally polyphasic (Moore et al. 2010), an approach which integrates phenotypic, genotypic and phylogenetic information (Vandamme et al. 1996). The recently published minimal standards for describing new species belonging to *Arcobacter, Campylobacter, Helicobacter* and *Wolinella* recommend a polyphasic approach which includes many of the same factors as included in previous minimal standards for the family *Campylobacteraceae* (Ursing, Lior, and Owen 1994) and the genus *Helicobacter* (Dewhirst, Fox, and On 2000) and proposes that sequence information from 16S rRNA, and additional phylogenetic markers, be used to support the position of the new taxon (On et al. 2017). The new standards also provide scope to include the whole genome sequence-based *in silico* analyses average nucleotide identity (ANI) and genome BLAST distance phylogeny (GBDP), in place of DNA-DNA hybridisation, to determine interspecific genomic relatedness (On et al. 2017).

Campylobacter concisus has small (0.5 x 4 μm) curved cells that exhibit rapid darting motility by using a single polar flagellum (Vandamme et al. 2005b). This species does not grow

[32] Also known as genomospecies

in microaerophilic atmospheres without hydrogen and is phenotypically heterogeneous (Vandamme et al. 2005b).

C. concisus has been isolated from a variety of sites in the human body including the gingival crevices of patients with gingivitis and periodontitis (Macuch and Tanner 2000); oesophagus and intestinal biopsies (Blackett et al. 2013, Hansen et al. 2013, Kaakoush et al. 2011, Mahendran et al. 2011, Zhang et al. 2009); blood (Vandamme et al. 1989); and a brain abscess (de Vries, Arents, and Manson 2008). In addition, *C. concisus* has been isolated from gastroenteritis-associated stool samples (Nielsen, Engberg, et al. 2013a) as well as those of healthy people (Engberg et al. 2000, Van Etterijck et al. 1996) leading to debate as to the role of this species in human gastroenteritis.

Culture independent methods, based on antigen detection or the polymerase chain reaction (PCR), are increasingly used to diagnose campylobacteriosis in humans (M'Ikanatha N et al. 2012, Zhang, Morrison, and Tang 2015). There are currently no published antigen-based methods for the detection of *C. concisus*. PCR for the identification of *C. concisus* have been described that target the 16S rRNA (Man et al. 2010, Marshall et al. 1999), 23S rRNA (Bastyns et al. 1995), *cpn*60 (Chaban et al. 2009, Inglis, Boras, and Houde 2011) or *gyr*B (Matsheka, Lastovica, and Elisha 2001) genes. Four of these PCR have been applied to the detection of *C. concisus* in human or animal stool samples (Chaban et al. 2009, Chaban, Ngeleka, and Hill 2010, Collado et al. 2013, Ferreira et al. 2014, Huq, Gonis, and Istivan 2014, Inglis, Boras, and Houde 2011, Kaakoush, Castano-Rodriguez, et al. 2014, Man et al. 2010, Samie et al. 2007, Underwood et al. 2016) but this is not routinely performed in New Zealand or in many laboratories globally. As demonstrated for other enteric pathogens such as non-O157 Shiga toxin-producing *Escherichia coli* (Shea et al. 2017), these culture independent methods can help elucidate the clinical importance of bacterial taxa. This is likely to also be the case for Epsilonproteobacterial taxa, including *C. concisus*. Stools yielding a positive signal with these culture independent tests still need to be cultured to provide information on antimicrobial resistance and/or subtypes (Jones and Gerner-Smidt 2012, Marder et al. 2017).

Isolation of *C. concisus* has been demonstrated from an inline milk filter (Serraino et al. 2013), chicken meat and minced beef (Lynch et al. 2011), and porcine caeca, carcasses and meat (Scanlon et al. 2013). In addition, this species has also been detected in chicken faeces (Kaakoush, Sodhi, et al. 2014), dog faeces (Chaban, Ngeleka, and Hill 2010) and saliva from cats with oral disease (Petersen et al. 2007). These isolations and detections demonstrate that *C. concisus* may be present in animals and birds but the low rates of isolation/detection observed for non-human sources suggest that humans are the natural host with the oral cavity being the primary colonisation site (Zhang et al. 2014).

The genomospecies concept, which indicates genetically distinct species without the necessary phenotypic marker for differentiation (Matsheka et al. 2002), has been proposed for *C. concisus* based on DNA-DNA hybridisation (Matsheka et al. 2002, Vandamme et al. 1989), pulsed field gel electrophoresis (PFGE, (Matsheka et al. 2002)), 23S rRNA polymerase chain reaction (PCR, (Engberg et al. 2005, Istivan et al. 2004, Kalischuk and Inglis 2011, On, Siemer, et al. 2013, Wang et al. 2017)), 23S rRNA gene sequence analysis (Chung et al. 2016), amplified fragment length polymorphism (AFLP) analysis, and multi-locus sequence typing (MLST) using either six (Chung et al. 2016, Ismail et al. 2012) or seven (Miller et al. 2012) housekeeping genes. A phylogenetic tree for 36 *C. concisus* based on the core-genome identified using the pan-genomic analysis tool Roary (Page et al. 2015), also identified two genomospecies (Chung et al. 2016).

The association of *C. concisus* with both healthy and diseased human tissue and the genetic heterogeneity observed for this species has led some researchers to hypothesise that there may be variation in the pathogenic potential of different *C. concisus* strains (Aabenhus et al. 2005, Deshpande et al. 2013, Kalischuk and Inglis 2011). A recent report suggests that *C. concisus* genomospecies 2 (GS2) is better adapted to the human gastrointestinal tract than genomospecies 1 (GS1) (Wang et al. 2017). There have been reports that suggest genes that might be involved in pathogenesis (Chung et al. 2016, Deshpande et al. 2013, Istivan et al. 2004, Kaakoush et al. 2011, Kalischuk and Inglis 2011, Mahendran et al. 2013) but there is

currently no consensus on what genes might be associated with differences in pathogenic potential.

A number of analyses are now available to evaluate the genetic relatedness of genomes based on nucleotide sequence including average nucleotide identity (ANI), tetranucleotide frequency (Tetra), genome BLAST distance phylogeny (GBDP), feature frequency profiling (FFP) and ribosomal multi-locus sequence typing (rMLST). ANI compares nucleotide utilisation between pairs of genomes by identifying orthologous genes using either basic local alignment search tool (BLAST) (Konstantinidis, Ramette, and Tiedje 2006, Konstantinidis and Tiedje 2005) or maximal unique matching (MUMmer) (Richter and Rossello-Mora 2009). Tetra was first reported by Pride et al. (2003) and is a fast, alignment-free method (Richter and Rossello-Mora 2009) that compares the frequency of the 256 different 4-nucleotide-words (tetramers) observed in the both strands of a genome's DNA against the expected frequencies, converts the differences to z-scores and compares the results for pairs of genomes (Teeling et al. 2004). GBDP uses high-scoring segment pairs (HSPs) or maximally unique matches (MUMs) to infer intergenomic distances for species delimitation (Auch, Klenk, and Goker 2010) as a form of digital DNA-DNA hybridisation (dDDH) (Auch et al. 2010). FFP is an alignment-free method that establishes the optimal nucleotide-word length for a set of genomes and then compares the frequencies of nucleotide-words of this length within the test genomes (Sims et al. 2009). rMLST is a form of MLST that aims to bridge the gap between bacterial subtyping and genealogy to the domain level by indexing the sequence variation in 53 ribosomal protein subunits (Jolley et al. 2012).

In addition to these published methods, it is also possible to perform comparisons based on amino acid sequence. An example of this involves clustering the core genomes using OrthoMCL which clusters orthologous genes using the Markov Cluster algorithm (Li, Stoeckert, and Roos 2003).

Comparisons at the functional level are possible using the carefully manually curated Clusters of Orthologous Groups of proteins (COG) database[33] which allows predictions of gene functions to be made based on the amino acid sequence similarity between unknown genes and genes that have been well studied experimentally (Tatusov, Koonin, and Lipman 1997). The database currently contains 4632 COGs classified into 26 functional categories (Galperin et al. 2015).

These tools were used to investigate the phylogenetic relationship between genomes from 22 *C. concisus* strains with the aim of better understanding the genetic diversity of this species.

3.3 Materials and Methods

3.3.1 Whole Genome Sequences

Eight well characterised *C. concisus* strains that have previously been reported to be genetically diverse (On, Siemer, et al. 2013) were grown on Columbia Horse Blood agar (Fort Richard, Auckland, New Zealand) at 37°C in a MAC-VA 500 cabinet (Don Whitely Scientific, Shipley, United Kingdom) with an atmosphere of 10% CO_2, 7% H_2, 3% O_2 and 80% N_2. *C. concisus* are challenging to grow in liquid media and are easily overgrown by contaminants so suspensions rather than liquid cultures were used for DNA extraction. Slightly turbid suspensions were prepared in 1 mL of Phosphate Buffered Saline (PBS; BR0014G, Oxoid, Basingstoke, England) by using a cotton-tipped applicator or disposable loop to transfer the colonies to the PBS. DNA was extracted using the Blood and Tissue Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions with RNaseI treatment included. The DNA quantity and quality was estimated using a Nanodrop 1000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA) and a more accurate quantity established using a Qubit (Life Technologies, Carlsbad, USA). The quality of the extracts were also checked by running on a

[33] http://www.ncbi.nlm.nih.gov/COG/

2% agarose gel in 1 X Tris-borate-EDTA (TBE) for 70 min at 110 V and visualising using ethidium bromide. At least 2 µg of DNA, in solutions of greater than 15 ng/µL, were produced for each strain. DNA libraries with ~550 bp inserts were prepared for each strain using the TruSeq™ DNA PCR-free Library Preparation (Illumina, San Diego, USA). The libraries were pooled with equal molarity before sequencing was undertaken on a MiSeq (Illumina). Velvet (version 1.2.10, (Zerbino and Birney 2008)) was used to assemble the short reads which were quality trimmed using SolexaQA++ (Cox, Peterson, and Biggs 2010) at a quality threshold of 0.01 (inferred base accuracy of 99% where 1 base in 100 is incorrect, or quality score of Q20), and then sorted by length to remove all resulting reads less than 50 bases long. Assemblies were performed in a paired end mode with scaffolding, using the additional single sequences that passed SolexaQA++ filtering. The output contigs were set to be at least 200 bp in length. To ensure that the best set of parameters for the assembly was used, all possible combinations of 20 kmer values (55 to 245 in increments of 10) and four random subsampling values (750,000 to 1,200,000 in steps of 150,000) were chosen to generate 80 assemblies per isolate. The "best" assembly was chosen by ranking the N_{50}, maximum contig length, overall bases, and number of contigs from each of the 80 assemblies per isolate, summing those ranks with equal weighting and picking the lowest resulting rank. N_{50} is the size of the last contig to bring the "combined contig size" to over half the size of all contigs when contigs are added to the "combined contig size" from largest to smallest (El-Metwally et al. 2013). In essence, 50% of the assembled bases are in a contig of a size equal or greater to the N_{50} score. The four metrics as well as the assembly ranking scores were calculated using in-house scripts.

One complete and seven draft (Deshpande et al. 2011, Deshpande et al. 2013) *C. concisus* genomes were publicly available from GenBank[34] on 24th December 2014. Another 234 genomes (47 complete, 187 draft) representing other *Campylobacter* species were also available from GenBank at this time. One GenBank genome was included in the study from

[34] http://www.ncbi.nlm.nih.gov/genbank

each of the following: the type species *C. fetus*, the clinically important *C. jejuni,* a selection of phylogenetically closely related species (*C. curvus, C. rectus, C. showae* and *C. mucosalis*) and two more distant species (*C. hominis* and *C. sputorum*) (see Figure 2). Genomes from type strains were used, where possible. Complete genomes for the type strains of *C. concisus* (since submitted to GenBank)*, C. rectus* and *C. showae* as well as one isolate each from *C. mucosalis* and *C. sputorum* were kindly provided by Dr William Miller (USDA). Four *C. concisus* genomes were also kindly shared before publication by Dr Mohsina Huq (RMIT University). An additional genome, 10_1_50, was only identified as *Campylobacter* sp. in GenBank was also included because it was tentatively identified as *C. concisus* in Chapter 4: Multiplex Ligation-dependent Probe Amplification on the basis of sequence similarity to PCR primers and taxon-specific genes.

All of the *C. concisus* genomes were obtained from strains isolated from humans. The clinical symptoms, demographics, geographical location of the host along with the year of isolation of the strains, are summarised in Table 2. Source information for the 10 genomes representing closely related species are included in Table 3.

Table 2: Characteristics of the 21 *C. concisus* Strains Used in the Comparative Genomics

Strain	Clinical details	Country	Site	Year[a]	GS[b]	Accession	Reference
10_1_50	Unknown	Unknown	Unknown	Unknown	2	ACWJ01000000	Unpublished
13826	Bloody stool & asthma	Denmark	Stool	1998	2	CP000792-4	(Aabenhus et al. 2005)
ATCC 33237^{T}	Gingivitis	USA	Oral cavity	1974	1	CP012541	(Aabenhus et al. 2005, Cornelius et al. 2017, Vandamme et al. 1989)
ATCC 51561	Healthy	Sweden	Stool	Unknown	2	ANNH01000000	(Deshpande et al. 2013)
ATCC 51562	Acute gastroenteritis	UK	Stool	Unknown	1	ANNI01000000	(Deshpande et al. 2013)
AUS22-Bd2	Irritable bowel syndrome	Australia	Duodenal biopsy	2013	1	LVLC01000000[d]	Unpublished
CCUG 19995	Pyrexia & exanthema	Sweden	Stool	1987	2	NDYN01000000	(Aabenhus et al. 2005, Vandamme et al. 1989, Cornelius et al. 2017)
Lasto28.99	Dysentery	South Africa	Stool	1999	1	NDYO01000000	(On, Siemer, et al. 2013, Cornelius et al. 2017)
Lasto61.99	Dysentery[e]	South Africa	Stool	1999	1	NEFM01000000	(On, Siemer, et al. 2013, Cornelius et al. 2017)
Lasto64.99	Dysentery	South Africa	Stool	1999	1	NDYP01000000	(On, Siemer, et al. 2013, Cornelius et al. 2017)

Strain	Clinical details	Country	Site	Year[a]	GS[b]	Accession	Reference
Lasto127.99	Chronic diarrhoea	South Africa	Stool	1999	2	NDYQ01000000	(On, Siemer, et al. 2013, Cornelius et al. 2017)
Lasto205.94	Bloody diarrhoea	South Africa	Stool	1994	1	NDYR01000000	(On, Siemer, et al. 2013, Cornelius et al. 2017)
Lasto220.96	Dysentery	South Africa	Stool	1996	1	NDYS01000000	(On, Siemer, et al. 2013, Cornelius et al. 2017)
Lasto393.96	Loose mucoid stools	South Africa	Stool	1996	1	NDYT01000000	(On, Siemer, et al. 2013, Cornelius et al. 2017)
RCH26	Gastroenteritis	Australia	Stool	2008	1	LVWL01000000, CM07854-5[d]	Unpublished
RMIT-JF1	Crohn's disease	Australia	Oral cavity	2013	1	JXUP01000000[d]	Unpublished
RMIT-O17	Healthy	Australia	Oral cavity	2014	1	LVLC01000000[d]	Unpublished
UNSW1	Chronic gastroenteritis	Australia	Intestinal biopsy	Unknown	2	ANNF01000000	(Deshpande et al. 2013, Kaakoush et al. 2011)
UNSW2	Crohn's disease	Australia	Intestinal biopsy	Unknown	2	ANNJ01000000	(Deshpande et al. 2013, Kaakoush et al. 2011)

Strain	Clinical details	Country	Site	Year[a]	GS[b]	Accession	Reference
UNSW3	Crohn's disease	Australia	Intestinal biopsy	Unknown	2	ANNE01000000	(Deshpande et al. 2013, Kaakoush et al. 2011)
UNSWCD	Crohn's disease	Australia	Intestinal biopsy	Unknown	2	AENQ01000000	(Deshpande et al. 2011)
UNSWCS	Acute gastroenteritis	Australia	Stool	Unknown	2	ANNG01000000	(Deshpande et al. 2013, Kaakoush et al. 2011)

[a] Year of isolation; [b] Genomospecies; [c] NCBI identification is *Campylobacter* sp., genomic analysis suggests identification of *C. concisus*; [d] Genomes kindly shared before publication by Dr Mohsina Huq, RMIT [e] *Shigella dysenteriae* co-infection

Table 3: Characteristics of the nine *Campylobacter* Strains Used in the Comparative Genomics

Strain	Species	Source	Site	Country	Year[a]	Accession	Reference
82-40	*C. fetus*	Unknown	Unknown	Unknown	Unknown	CP000487	Unpublished
ATCC 33238^T	*C. rectus*	Human	Periodontal pocket	USA	1974	Unpublished[b]	(Tanner et al. 1981)
ATCC 33560^T	*C. jejuni*	Bovine	Faeces	Belgium	Unknown	AIOL01000000	Unpublished
ATCC 51146^T	*C. showae*	Human	Gingival crevice	USA	1992	Unpublished	(Etoh et al. 1993)
ATCC BAA-381^T	*C. hominis*	Human	Stool	UK	1998	CP000775-6	(Lawson et al. 2001)
CCUG 20703	*C. sputorum*	Bovine	Faeces	UK	Unknown	Unpublished	(Vandamme et al. 1990)
CCUG 21559	*C. mucosalis*	Porcine	Necrotic colitis	Unknown	Unknown	Unpublished	(Korczak et al. 2006, Lawson et al. 1981)
DSM 6644^T	*C. curvus*	Human	Jaw abscess	USA	Unknown	AQXN01000000	Unpublished
DSM 21682	*C. mucosalis*	Porcine	Small intestine	UK	1974	JHQQ01000000	Unpublished

[a] Year of isolation; [b] Unpublished genomes kindly shared by Dr William Miller, USDA

All of the genomes were saved in Geneious[35] (R6.1.7 and R9.1.7) and fasta files of the nucleotide sequences were exported. The contigs of draft genomes were concatenated to form a single document and joined with the addition of 200 N's between contigs to produce a single sequence per genome. All of the 31 genomes, including the 14 publicly available genomes, the five genomes kindly provided by Dr William Miller (USDA), the four genomes kindly provided by Dr Mohsina Huq (RMIT University) and the 8 *C. concisus* genomes sequenced in this project, were annotated using Prokka (version 1.8; (Seemann 2014)) to provide a consistent annotation procedure for comparison.

3.3.2 Comparative Genomic Analysis

ANI, using both BLAST (ANIb) and MUMmer (ANIm), and Tetra were undertaken using the jSpecies V1.2.1 website[36]. GBDP was performed using the Genome-to-Genome Distance Calculator (GGDC) 2.0 available online[37] using the recommended BLAST+. Heatmaps were generated for the ANI, Tetra and GBDP data in R (v 3.2.5, (R Core Team 2016)). FFP (version 3.19) was downloaded[38] and the optimal *l*-mer established using supplied scripts and as described by (Sims et al. 2009). The suggested pipeline was used to produce a matrix. rMLST analysis was undertaken using a custom script that first performed a BLAST search using a reference set of *Campylobacter* genus ribosomal protein genes to identify target genes in Prokka-annotated genomes. The rMLST genes were then extracted and the sequences submitted to the rMLST database on the pubMLST website[39], from which allele names were returned in tabular form. OrthoMCL (version 2.0.9) (Li, Stoeckert, and Roos 2003) and custom Perl scripts were used on the amino acid sequences of the genomes to generate clusters of

[35] http://www.geneious.com/

[36] http://imedea.uib-csic.es/jspecies/download.html

[37] http://ggdc.dsmz.de/distcalc2.php

[38] http://sourceforge.net/projects/ffp-phylogeny/

[39] http://pubmlst.org/rmlst/

orthologous genes (COG) as described by Grange et al. (2016). Genes in a cluster were aligned individually, then concatenated as gapped alignments to preserve logical gene alignments. A custom Perl script written by Dr Patrick Biggs was performed three types of Position-Specific Iterative-BLAST (PSI-BLAST) searches on the amino sequence of all protein encoding genes for each genome against a local copy of the COGs database[40], as per COGnitor software[41] and using the default values. COGmakehashLSCN, COGreadblast and COGnitor were then run with default parameters. A summary of the number of paralogues of each COG observed for each genome was generated in a MySQL database. This was then converted to a distance matrix using a custom Perl script which tabulated the COG allocations of the predicted genes from each genome against a master list of all COG genes providing a count per genome of each COG. The COGs were then labelled as present, absent or variable for the set of genomes and variable COGs extracted as a matrix and converted to a distance matrix using either Euclidean or Manhattan distances and a standard distance matrix equation. The matrix outputs of FFP, rMLST, OrthoMCL and COGs were converted to NEXUS format (Maddison, Swofford, and Maddison 1997) and visualised using SplitsTree4 (version 4.12.6, (Huson and Bryant 2006)).

The 31 genomes were evaluated using the genome assembly quality assessment tool QUAST (Gurevich et al. 2013) using the online calculator[42]. A GC content boxplot was generated for the 22 *C. concisus* genomes using R (v 3.2.5, (R Core Team 2016)).

16S rRNA gene sequence analysis was undertaken in Geneious (R9.1.7). The “Extract Annotations” tool extracted Prokka-annotated 16S rRNA genes from all 31 genomes. For two *C. concisus* genomes, Lasto205.94 and UNSW2, the 16S rRNA genes were fragmented on different contigs within the WGS. A BLAST search of the *C. jejuni* ATCC 33560^T 16S rRNA gene was used to identify the fragments in these two genomes which were then aligned using

[40] http://www.ncbi.nlm.nih.gov/COG/, 2014 update

[41] https://www.ncbi.nlm.nih.gov/Web/Newsltr/Winter01/index.html

[42] Available at http://quast.bioinf.spbau.ru/

the ATCC 33560^T as the reference. Two 16S rRNA constructs were generated for each of these two genomes. Phylogenetic trees were generated in Geneious (R9.1.7) with the Geneious Tree Builder tool using the Tamura–Nei genetic distance model and neighbor-joining with a 65% similarity cost matrix, a gap open penalty of 12, a gap extension penalty of 3, and global alignment with free end gaps as the alignment type. A Clustal W (Thompson, Higgins, and Gibson 1994) alignment was also generated in Geneious (R9.1.7) for the 33 16S rRNA gene sequences and the percentage identity matrix exported as comma separated values (csv) so the data could be evaluated in Excel (2013, Microsoft, Redwood, WA).

3.3.3 Identification of Possible Genomospecies-Specific Genes

Large-scale BLAST score ratio (LS-BSR, (Sahl et al. 2014)) was used to evaluate the pan-genome of *C. concisus* and the `listAandB.pl` script written by Dr Patrick Biggs was used to identify coding sequences (CDS, called centroids by LS-BSR) present in one genomospecies and absent from the other (see section 4.3.2 Pan-genomic Analysis and Target Gene Selection). The `listAandB.pl` script was used in conjunction with two lists of genomes (listA and listB) for which BSR criteria (e.g. >0.8 or <0.4) were defined. The LS-BSR output matrix was filtered on the basis of criterion A for genomes in listA and criterion B for genomes in listB and a list containing the CDS that fulfilled both criteria was generated. The locus_tag and annotation for GS1- and GS2-specific CDS were retrieved from the GenBank versions of the ATCC 33237^T and 13826 genomes, respectively. The genomospecies-specific CDS identified from the *C. concisus* LS-BSR were compared to those identified using Roary (Page et al. 2015) in Chung et al. (2016) and the specificity of the CDS were further evaluated by searching the local EpsiloFsa, KrunoFsa and RMITConcisus BLAST databases generated in section 4.3.1 Epsilonproteobacterial Genomes using the CDS as queries, a word size of 11 and a maximum number of hits of 1,000, 10 and 10, respectively. Grade, a weighted sequence similarity metric calculated by Geneious from query coverage, e-value and identity (Biomatters Ltd 2015), was used to sort the results as this placed the longest hits with significant sequence similarity

towards the top rather then shorter hits with higher percentage identity. The COG numbers and functional groups were identified for the GS1- and GS2-specific CDS using eggNOG-mapper[43] (Huerta-Cepas et al. 2017).

3.4 Results

3.4.1 Whole Genome Sequences

The eight draft *C. concisus* genomes generated in this project and submitted to GenBank (Cornelius et al. 2017) consisted of between 13 and 44 contigs, with N_{50} values between 134,605 and 349,534 bp (Table 4).

Table 4: Characteristics of the Eight Draft *Campylobacter concisus* Genomes Generated in this Project

Strain	Accession	Genome size (bp)	Nº contigs[a]	N_{50} (bp)[a]	GC (%)[a]	GS
CCUG 19995	NDYN01000000	2,105,146	44	215,794	39.39	2
Lasto28.99	NDYO01000000	1,937,734	30	150,974	37.73	1
Lasto61.99	NEFM01000000	1,836,142	18	260,338	37.55	1
Lasto64.99	NDYP01000000	1,872,282	20	190,490	37.49	1
Lasto127.99	NDYQ01000000	2,027,572	44	134,605	39.40	2
Lasto205.94	NDYR01000000	1,891,545	39	179,898	37.62	1
Lasto220.96	NDYS01000000	1,796,488	13	349,534	37.48	1
Lasto393.96	NDYT01000000	1,854,434	15	256,243	37.36	1

[a] Calculated using the online version of QUAST (http://quast.bioinf.spbau.ru/)

3.4.2 Comparative Genomic Analysis

ANIb, ANIm and Tetra were undertaken on the 22 *C. concisus* genomes and nine related species using the jSpecies website. The heatmap generated for ANIb (Figure 8) demonstrates the separation of the *C. concisus* genomes into two genomospecies (GS) which are also well separated from all related *Campylobacter* spp. Only the two intra-*C. mucosalis* pairwise results exceeded 97% (Figure 8). All of the ANIb *C. concisus* intra-GS pairwise results were between

[43] http://eggnogdb.embl.de/#/app/emapper

93% and 96%, all of the *C. concisus* inter-GS and *C. rectus/C. showae* inter-species pairwise results were between 85% and 89% and all of the remaining inter-species pairwise results were <75%.

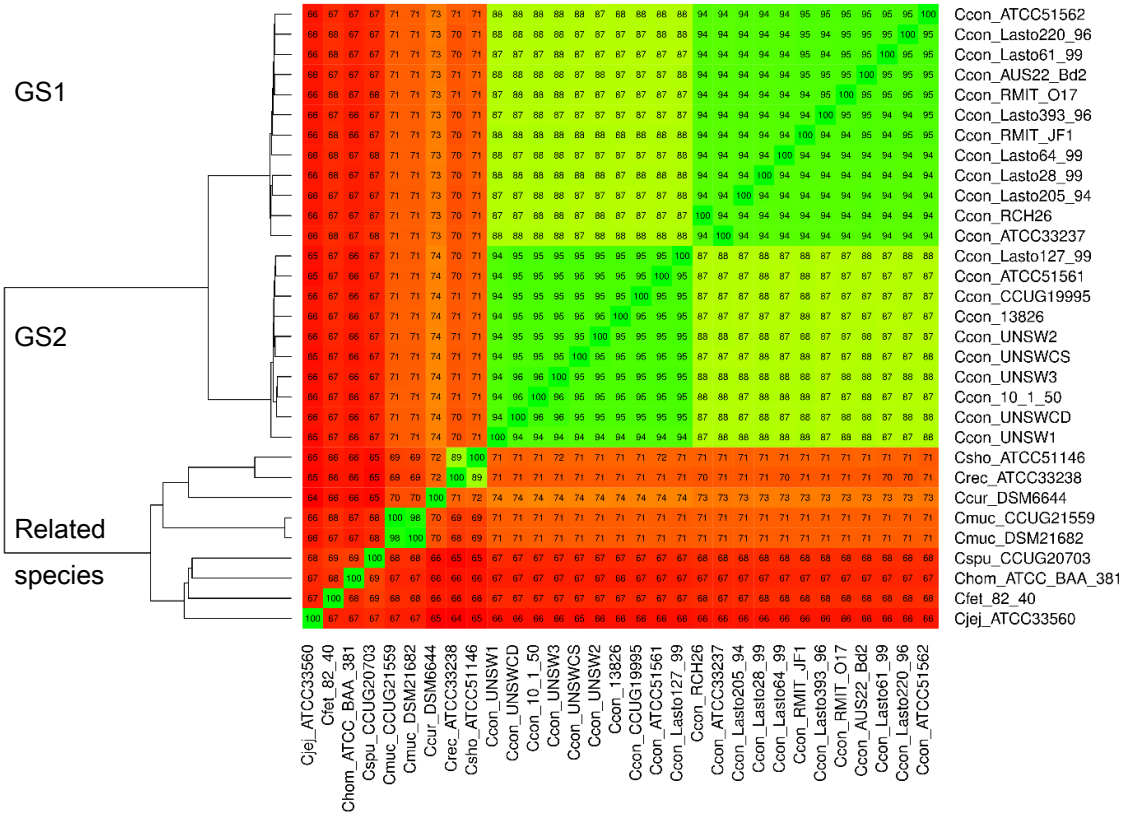


Figure 8: Heatmap for Average Nucleotide Identity (using BLAST) for 22 *C. concisus* Genomes and nine Genomes Representing Related Species

The results were rounded to whole numbers to aid readability.

Ccon *C. concisus*, Ccur *C. curvus*, Cfet *C. fetus*, Chom *C. hominis*, Cjej *C. jejuni*, Cmuc *C. mucosalis*, Crec *C. rectus*, Csho *C. showae*, Cspu *C. sputorum*

The heatmap for ANIm (Figure 9) shows the same two *C. concisus* genomospecies as those produced by ANIb. Only the two intra-*C. mucosalis* pairwise results exceeded 97%. All of the ANIm *C. concisus* intra-GS pairwise results were between 93% and 96%. Almost all (89/90, 98.9%) of the *C. concisus* GS2 and 6 of 122 (4.9%) GS1 intra-GS pairwise ANIm results exceeded 95%. All of the *C. concisus* inter-GS as well as the inter-species pairwise results were less than 91%.

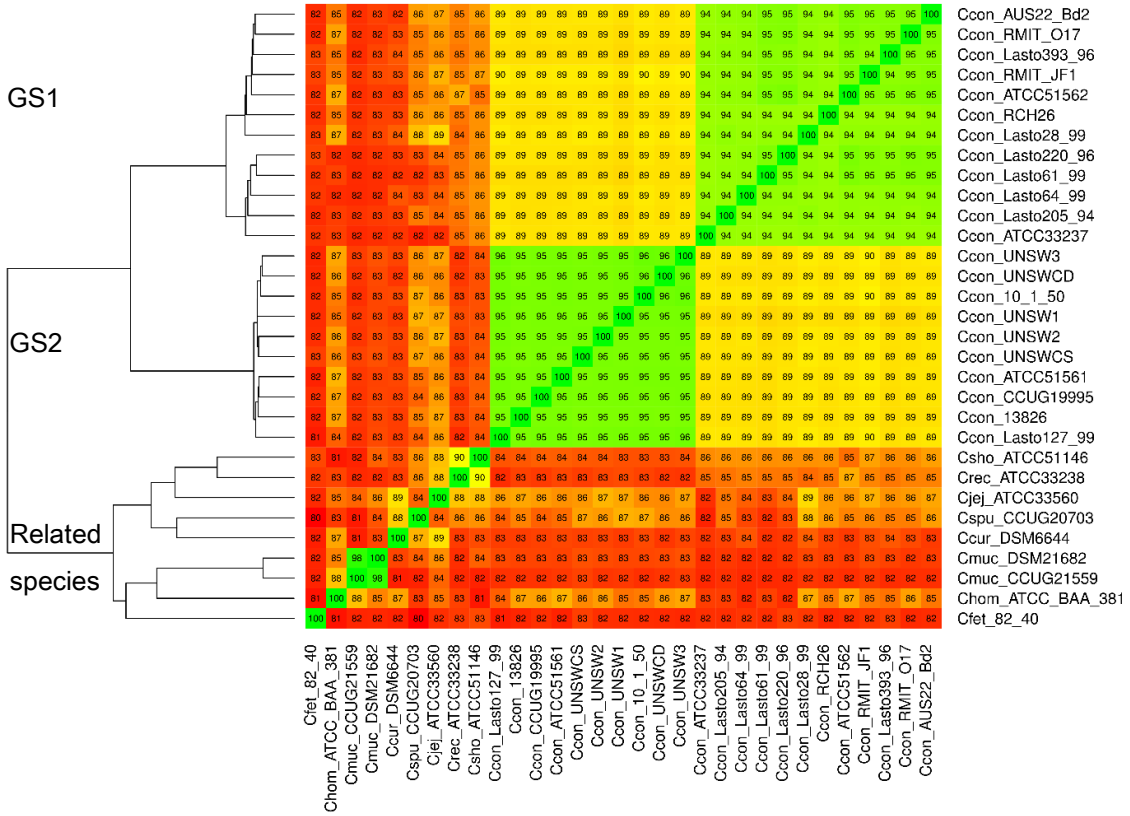


Figure 9: Heatmap for Average Nucleotide Identity (using MUMmer) for 22 *C. concisus* Genomes and nine Genomes Representing Related Species

The results were rounded to whole numbers to aid readability.

Ccon *C. concisus*, Ccur *C. curvus*, Cfet *C. fetus*, Chom *C. hominis*, Cjej *C. jejuni*, Cmuc *C. mucosalis*, Crec *C. rectus*, Csho *C. showae*, Cspu *C. sputorum*

The heatmap generated for Tetra results (Figure 10) separated the *C. concisus* genomes into the same two genomospecies as observed for ANI. The intra-*C. mucosalis* pairs and 70% of the *C. concisus* intra-GS pairs were at least 0.999. The remaining 30% of the *C. concisus* intra-GS pairs were 0.997 or 0.998. All of the *C. concisus* inter-GS and the *C. rectus/C. showae* inter-species pairwise results were greater than 0.98 and less than 0.99. All of the remaining inter-species pairwise results were <0.95.

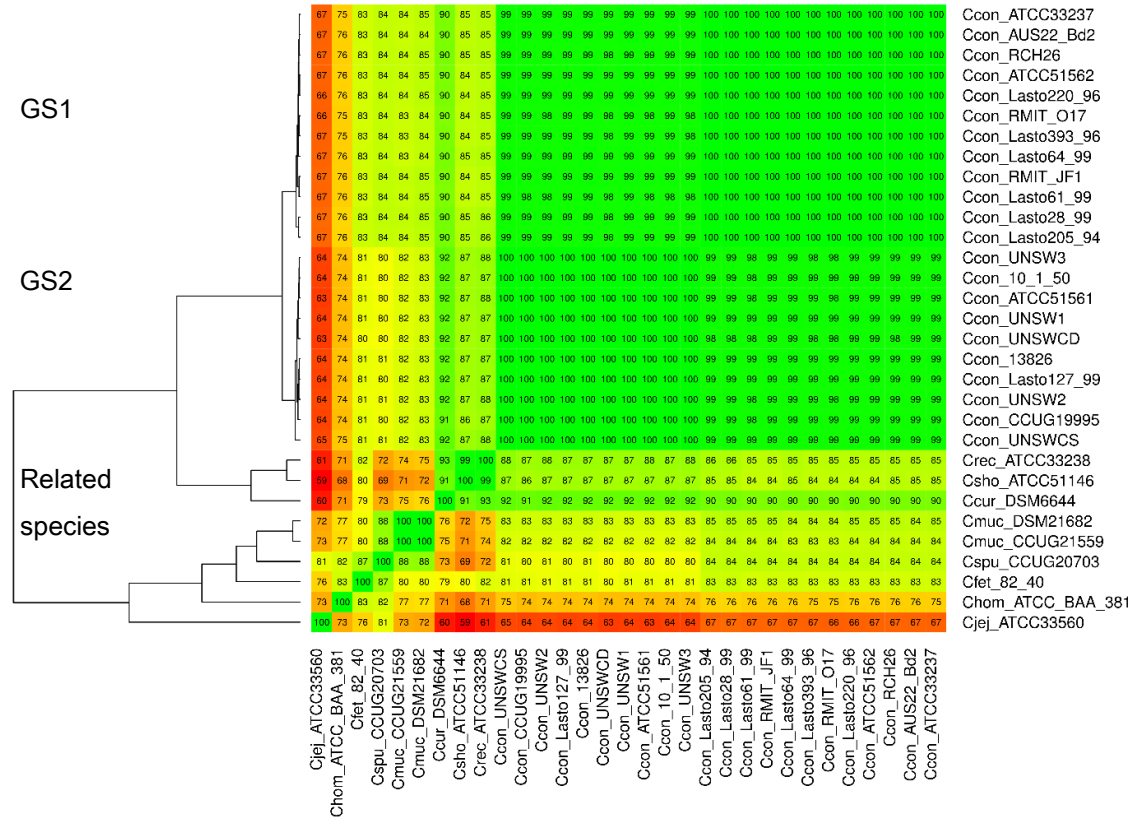
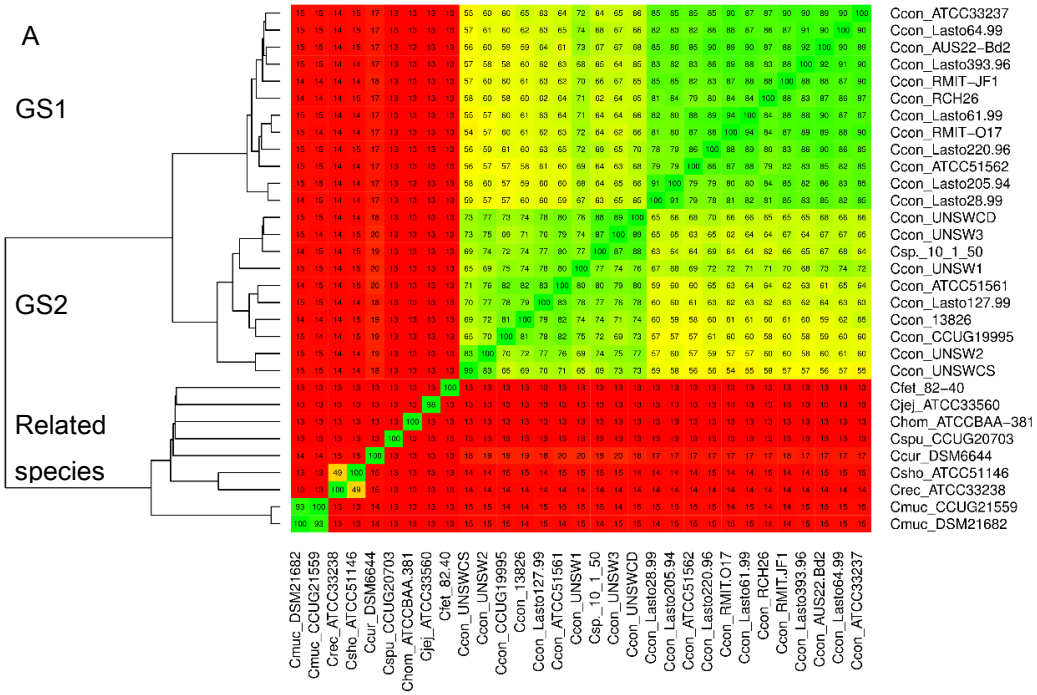


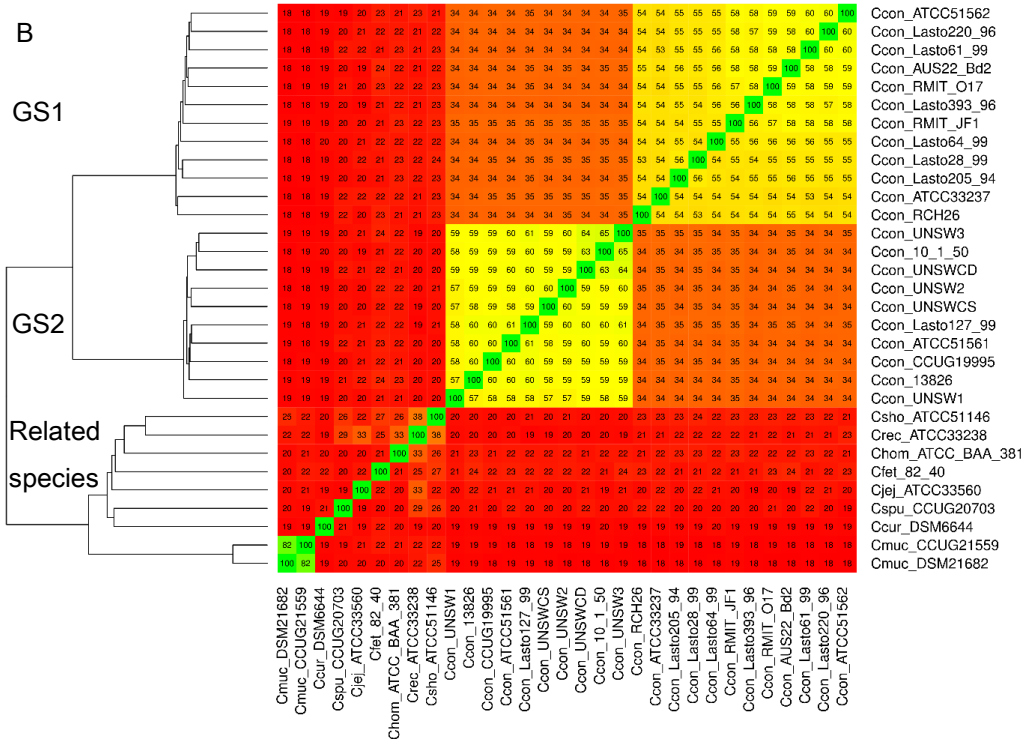
Figure 10: Heatmap for Tetranucleotide Frequency for 22 *C. concisus* Genoms and nine Genomes Representing Related Species

The results were rounded to whole numbers to aid readability.

Ccon *C. concisus*, Ccur *C. curvus*, Cfet *C. fetus*, Chom *C. hominis*, Cjej *C. jejuni*, Cmuc *C. mucosalis*, Crec *C. rectus*, Csho *C. showae*, Cspu *C. sputorum*

GBDP was performed on the 22 *C. concisus* genomes and nine genomes representing related species using GGDC and heatmaps were generated for each of the three formula (Figure 11). All three formulas separated the *C. concisus* genomes into the two genomospecies observed for ANI and Tetra. The *C. mucosalis* intra-species pairwise results were >70% for all formula. No *C. concisus* intra- or inter-GS pairwise results were >70% using formula 2. All of the *C. concisus* GS1 intra-GS pairwise results using formulas 1 and 3 were >70% but 14 of the 90 (15.6%) *C. concisus* GS2 intra-GS pairwise results were <70%. In addition, 14 of the 240 (5.8%) *C. concisus* inter-GS pairwise results were >70% using formula 1. Using formula 3, all of the *C. concisus* inter-GS pairwise results were between 48.9% and 63.2% and all of the inter-species GBDP pairwise results were <19% except between *C. rectus* and *C. showae* where the result was 46%.


A
GS1
GS2
Related species
Ccon_ATCC33237
Ccon_Lasto64.99
Ccon_AUS22-Bd2
Ccon_Lasto393.96
Ccon_RMIT-JF1
Ccon_RCH26
Ccon_Lasto61.99
Ccon_RMIT-O17
Ccon_Lasto220.96
Ccon_ATCC51562
Ccon_Lasto205.94
Ccon_Lasto28.99
Ccon_UNSWCD
Ccon_UNSW3
Csp._10_1_50
Ccon_UNSW1
Ccon_ATCC51561
Ccon_Lasto127.99
Ccon_13826
Ccon_CCUG19995
Ccon_UNSW2
Ccon_UNSWCS
Cfet_82-40
Cjej_ATCC33560
Chom_ATCCBAA-381
Cspu_CCUG20703
Ccur_DSM6644
Csho_ATCC51146
Crec_ATCC33238
Cmuc_CCUG21559
Cmuc_DSM21682
Cmuc_DSM21682
Cmuc_CCUG21559
Crec_ATCC33238
Csho_ATCC51146
Ccur_DSM6644
Cspu_CCUG20703
Chom_ATCCBAA_381
Cjej_ATCC33560
Cfet_82_40
Ccon_UNSW1
Ccon_13826
Ccon_CCUG19995
Ccon_ATCC51561
Ccon_Lasto127_99
Ccon_UNSWCS
Ccon_UNSW2
Ccon_UNSWCD
Ccon_10_1_50
Ccon_UNSW3
Ccon_RCH26
Ccon_ATCC33237
Ccon_Lasto205_94
Ccon_Lasto28_99
Ccon_Lasto64_99
Ccon_RMIT_JF1
Ccon_Lasto393_96
Ccon_RMIT_O17
Ccon_AUS22_Bd2
Ccon_Lasto61_99
Ccon_Lasto220_96
Ccon_ATCC51562


B
GS1
GS2
Related species
Ccon_ATCC51562
Ccon_Lasto220_96
Ccon_Lasto61_99
Ccon_AUS22_Bd2
Ccon_RMIT_O17
Ccon_Lasto393_96
Ccon_RMIT_JF1
Ccon_Lasto64_99
Ccon_Lasto28_99
Ccon_Lasto205_94
Ccon_ATCC33237
Ccon_RCH26
Ccon_UNSW3
Ccon_10_1_50
Ccon_UNSWCD
Ccon_UNSW2
Ccon_UNSWCS
Ccon_Lasto127_99
Ccon_ATCC51561
Ccon_CCUG19995
Ccon_13826
Ccon_UNSW1
Csho_ATCC51146
Crec_ATCC33238
Chom_ATCC_BAA_381
Cfet_82_40
Cjej_ATCC33560
Cspu_CCUG20703
Ccur_DSM6644
Cmuc_CCUG21559
Cmuc_DSM21682
Cmuc_DSM21682
Cmuc_CCUG21559
Ccur_DSM6644
Cspu_CCUG20703
Cjej_ATCC33560
Cfet_82_40
Chom_ATCC_BAA_381
Crec_ATCC33238
Csho_ATCC51146
Ccon_UNSW1
Ccon_13826
Ccon_CCUG19995
Ccon_ATCC51561
Ccon_Lasto127_99
Ccon_UNSWCS
Ccon_UNSW2
Ccon_UNSWCD
Ccon_10_1_50
Ccon_UNSW3
Ccon_RCH26
Ccon_ATCC33237
Ccon_Lasto205_94
Ccon_Lasto28_99
Ccon_Lasto64_99
Ccon_RMIT_JF1
Ccon_Lasto393_96
Ccon_RMIT_O17
Ccon_AUS22_Bd2
Ccon_Lasto61_99
Ccon_Lasto220_96
Ccon_ATCC51562

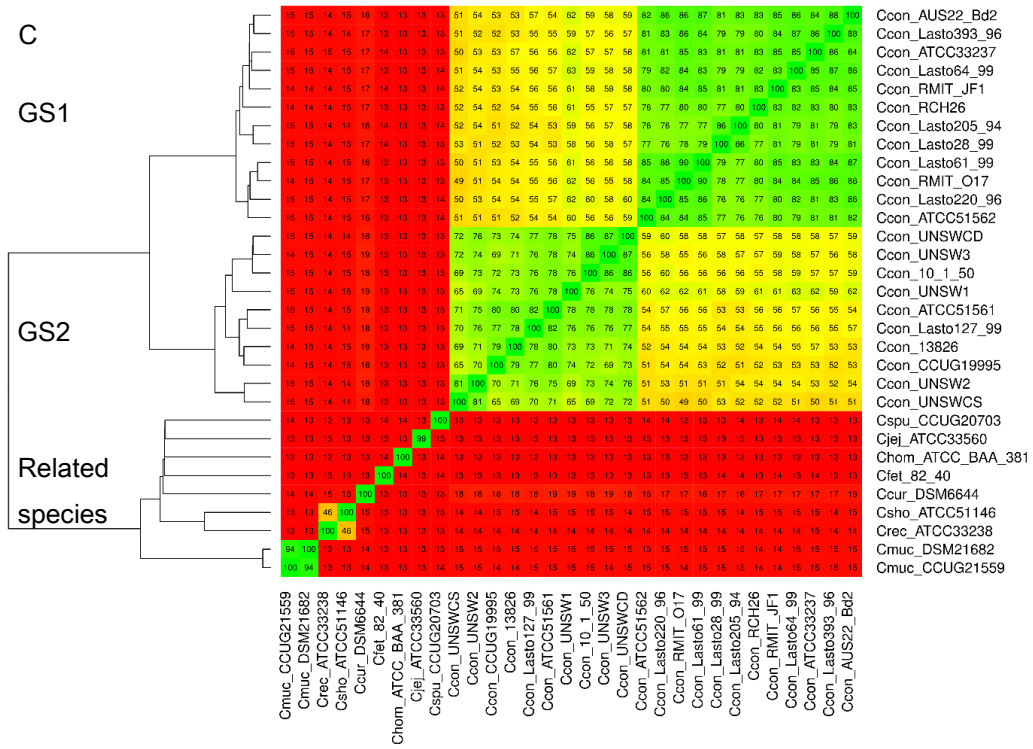


Figure 11: Heatmaps for Genome BLAST Distance Phylogeny (GBDP) for 22 *C. concisus* Genomes and nine Genomes Representing Related Species

Panel A. GBDP using Formula 1 HSP length/total length. Panel B. GBDP using Formula 2 identities/HSP length. Panel C. GBDP using Formula 3 identities/total length. The results were rounded to whole numbers to aid readability.
Ccon *C. concisus*, Ccur *C. curvus*, Cfet *C. fetus*, Chom *C. hominis*, Cjej *C. jejuni*, Cmuc *C. mucosalis*, Crec *C. rectus*, Csho *C. showae*, Cspu *C. sputorum*

FFP involves two steps (Sims et al. 2009). The upper limit of optimal *l*-mer length is established first by calculating the relative entropies for a range of *l*-mer lengths and identifying the length at which the relative entropy approaches 0. The second step involves evaluating the stability of the clusters within NeighborNets generated from FFPs of *l*-mer lengths smaller than, and equal to, the optimal length upper limit. Relative entropies for *l*-mer lengths 4 to 24 were generated to establish the upper limit of the optimal length (Figure 12). The maximum optimal *l*-mer length, with a relative entropy of 0.029732, was 20 nucleotides. The same three clearly separated clusters (two *C. concisus* genomospecies and related species) observed for the 31 genomes using ANI, Tetra and GBDP were also observed for FFP when the *l*-mer lengths were 11 to 20 nucleotides and more tree-like, rather than network-like, NeighborNets were observed for *l*-mer lengths of greater than 15 nucleotides (Appendix I) which indicates that the signal is

less conflicted (Huson and Bryant 2006). The NeighborNet generated using SplitsTree4 from the FFP results using an *l*-mer of length 20 (Figure 13) is typical of the results observed.

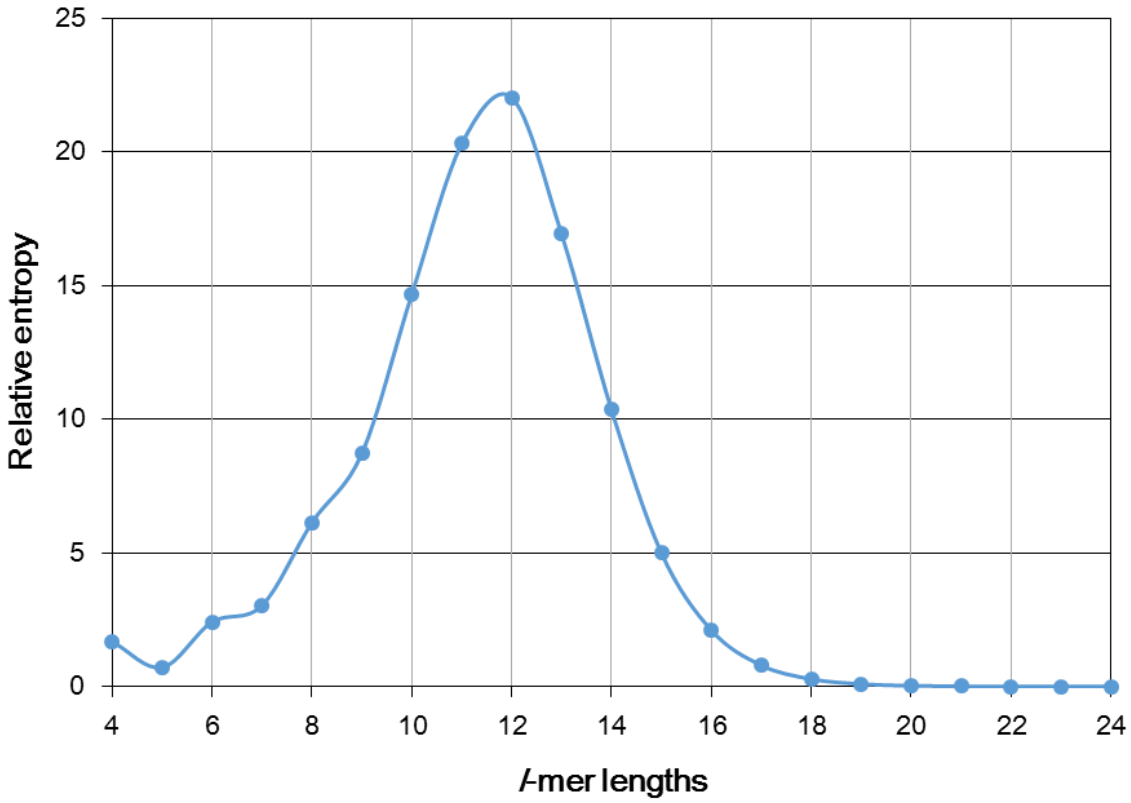


Figure 12: Graph of Relative Entropies for *l*-mer Lengths

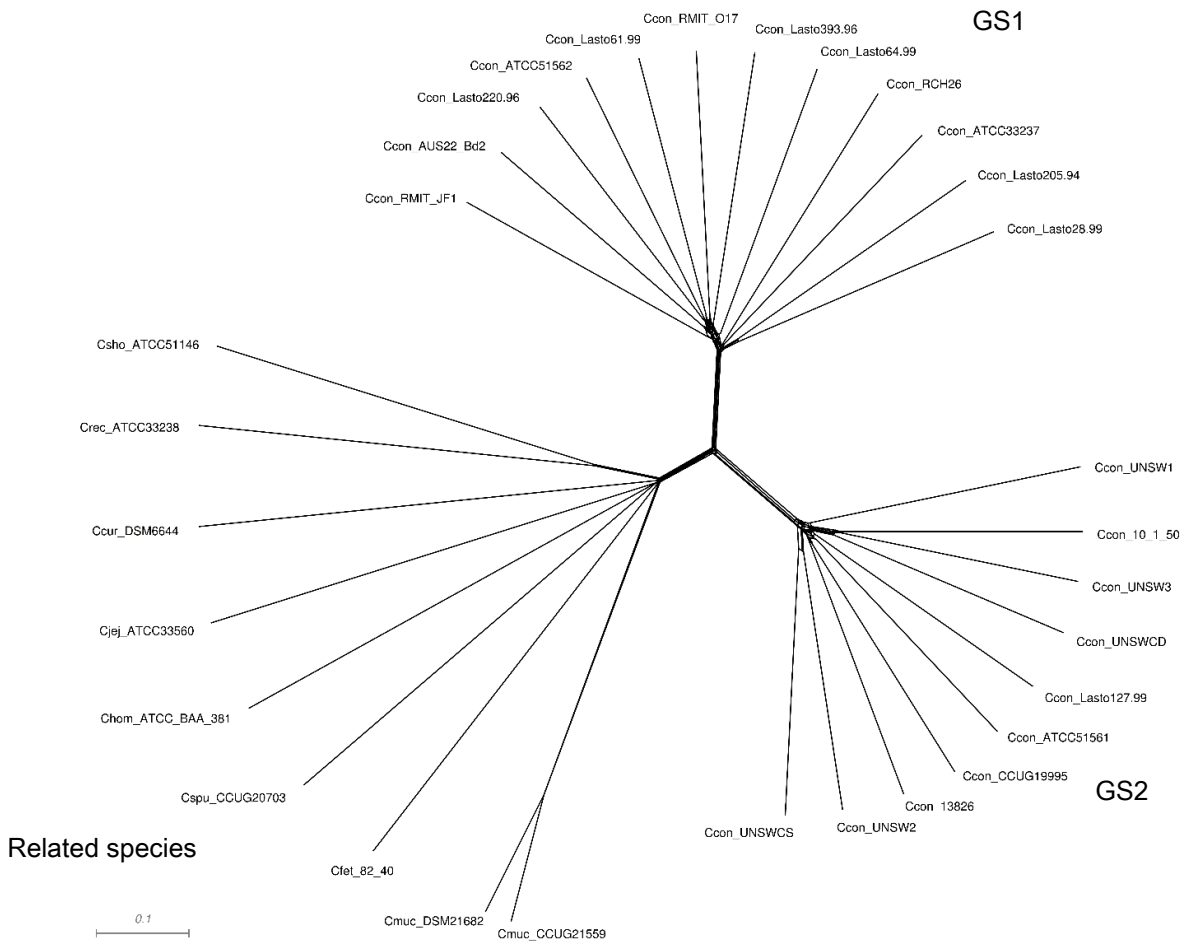


Figure 13: NeighborNet visualised in SplitsTree4 of Feature Frequency Profiling *l*-mer Length 20 Results for 22 *C. concisus* Genomes and nine Genomes Representing Related Species

Ccon *C. concisus*, Ccur *C. curvus*, Cfet *C. fetus*, Chom *C. hominis*, Cjej *C. jejuni*, Cmuc *C. mucosalis*, Crec *C. rectus*, Csho *C. showae*, Cspu *C. sputorum*

The rMLST alleles for the 31 genomes are summarised in Appendix II. Of the 21042 nucleotides included in the rMLST analysis, 8506 (40.4%) were variable. The separation of the two *C. concisus* genomospecies and related species were also observed in the NeighborNets generated from the rMLST nucleotide sequences using the 31 genomes (Figure 14).

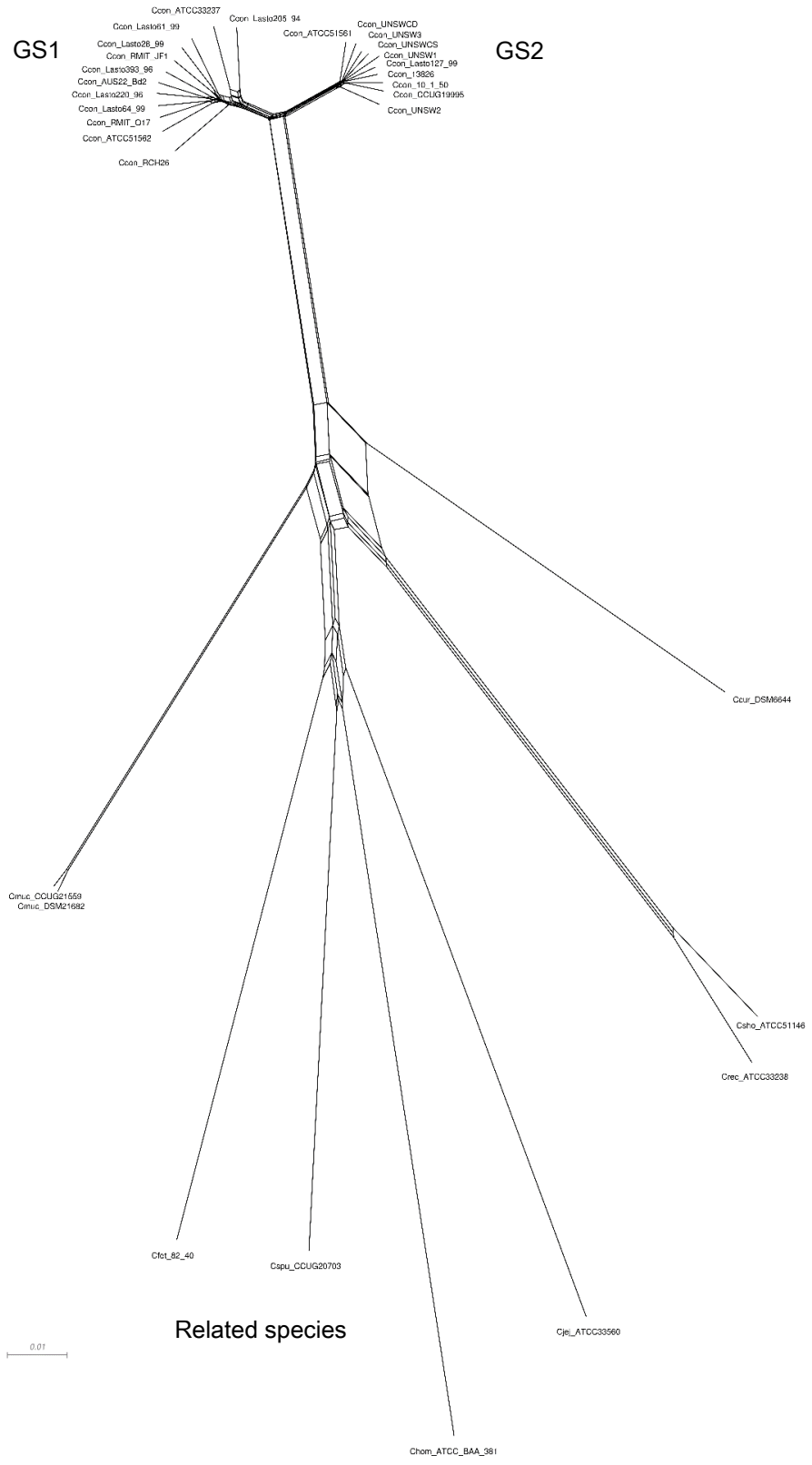


Figure 14: NeighborNet of Ribosomal Multi-Locus Sequence Typing of 22 *C. concisus* Genomes and nine Genomes Representing Related Species Using Nucleotide Sequence

Ccon *C. concisus*, Ccur *C. curvus*, Cfet *C. fetus*, Chom *C. hominis*, Cjej *C. jejuni*, Cmuc *C. mucosalis*, Crec *C. rectus*, Csho *C. showae*, Cspu *C. sputorum*

A total of 179,781 amino acids from the 31 genomes were included in the OrthoMCL analysis, with 115,185 (64.1%) sites being variable. The resulting clusters included 568 genes with a minimum length for inclusion in a cluster being at least 80% of the length of the longest ortholog. The separation of the two *C. concisus* genomospecies and related species were also observed in the NeighborNets generated from the amino acid-based OrthoMCL analyses (Figure 15) using the 31 genomes.

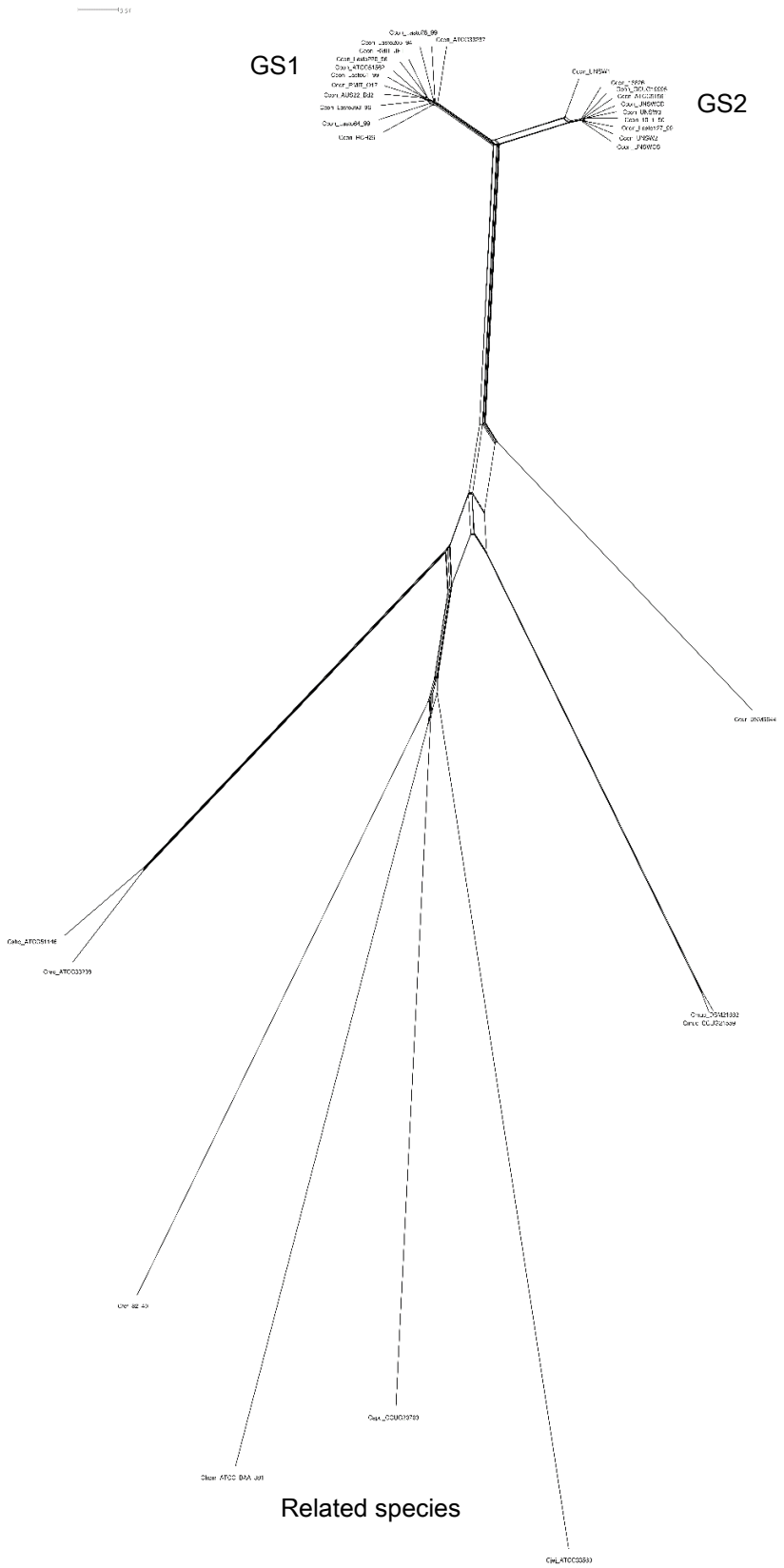


Figure 15: NeighborNet of OrthoMCL Results for 22 *C. concisus* Genomes and nine Genomes Representing Related Species using Amino Acid Sequences

Ccon *C. concisus*, Ccur *C. curvus*, Cfet *C. fetus*, Chom *C. hominis*, Cjej *C. jejuni*, Cmuc *C. mucosalis*, Crec *C. rectus*, Csho *C. showae*, Cspu *C. sputorum*

Supplementary File 1 summarises the number of genes for each COG functional group that were identified in the 31 genomes. The same two *C. concisus* genomospecies clusters separate from the related species were also observed in the NeighborNet generated from the COGs analysis which infers function based on amino acid sequence (Figure 16).

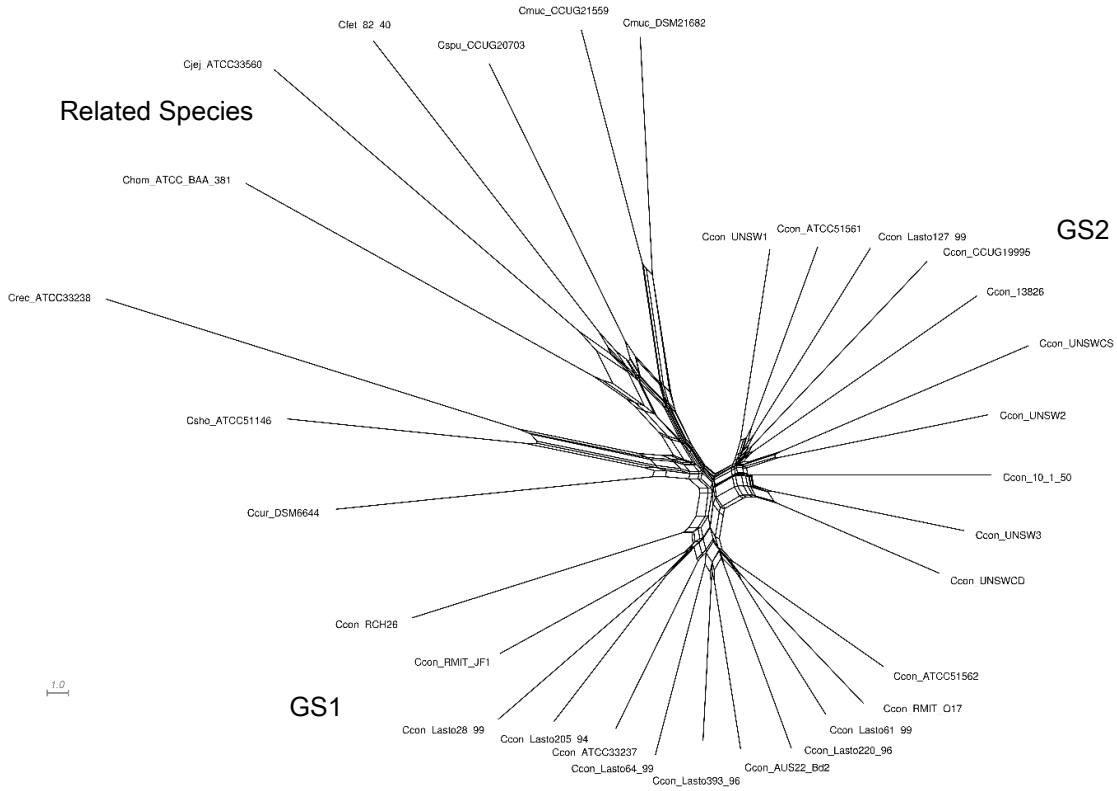


Figure 16: NeighborNet Visualised in SplitsTree4 Generated from Clusters of Orthologous Genes Analysis of 22 *C. concisus* Genomes and nine Genomes Representing Related Species

Ccon *C. concisus*, Ccur *C. curvus*, Cfet *C. fetus*, Chom *C. hominis*, Cjej *C. jejuni*, Cmuc *C. mucosalis*, Crec *C. rectus*, Csho *C. showae*, Cspu *C. sputorum*

The median (and approximate 95% confidence interval) GC contents for *C. concisus* GS1 and GS2 were 37.585% (37.487 – 37.683) and 39.455% (39.305 – 39.605), respectively (Figure 17). The GC contents of ATCC 33237^T and CCUG 19995 were 37.62 and 39.39, respectively. All of the *C. concisus* intra-GS GC difference results generated as part of the GBDP results from GGDC were <1 and all of the *C. concisus* inter-GS results were >1 (Figure 18). Interestingly, 10 of the 12 GC difference results for pairs involving the *C. mucosalis* genome DSM 21682^T (36.64 %GC) and *C. concisus* GS1 genomes were also <1, as were 5 of the 12 pairs involving the *C.*

mucosalis genome CCUG 21559 (36.54 %GC) and the *C. concisus* GS1 genomes. Two other inter-species pairs (*C. jejuni* / *C. sputorum* and *C. rectus* / *C. curvus*) also had GC difference results <1.

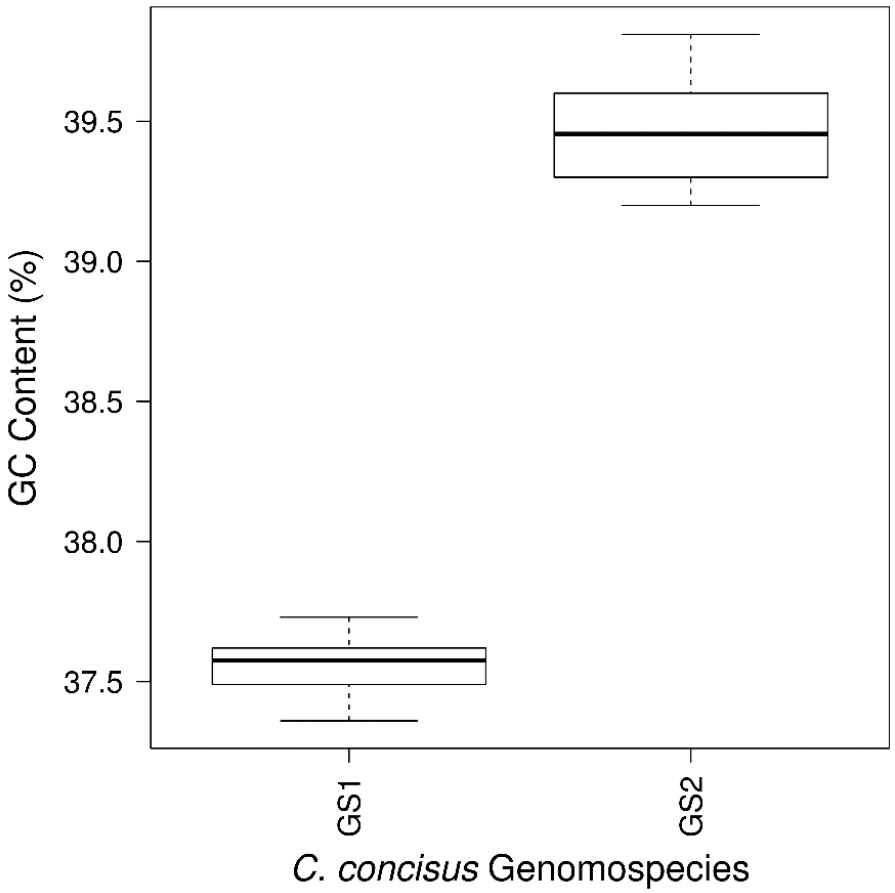


Figure 17: Boxplots of the GC Content of *C. concisus* Genomospecies GS1 and GS2

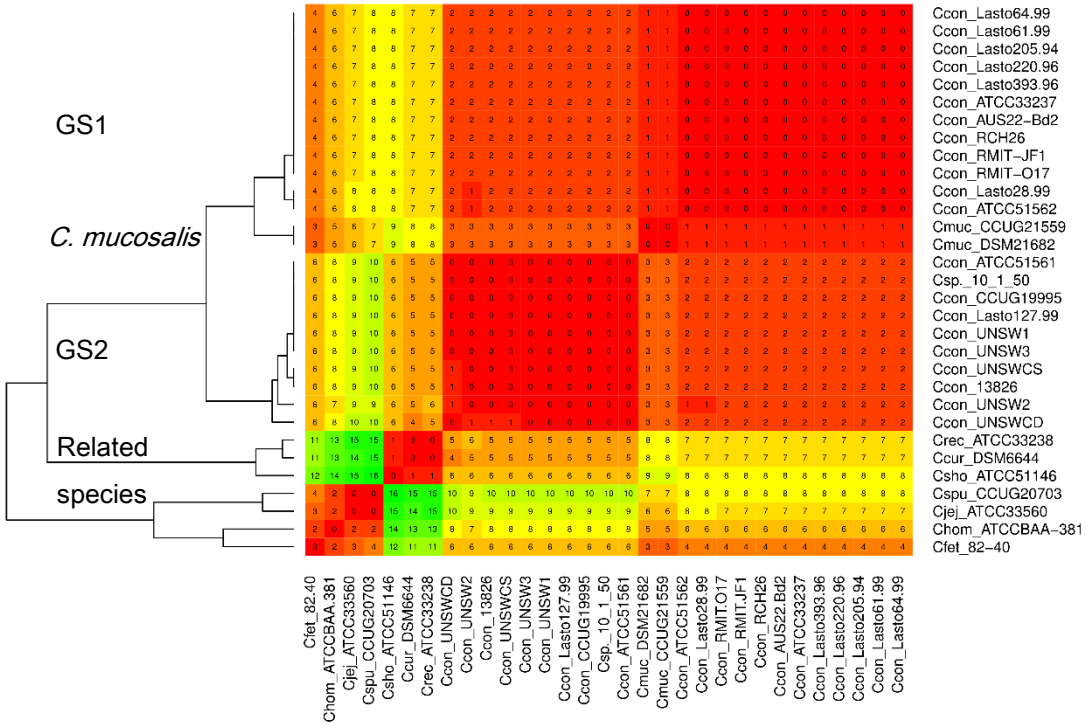


Figure 18: Heatmap of GC Differences for 22 *C. concisus* Genomes and nine Genomes Representing Related Species

Results are rounded to aid readability. Ccon *C. concisus*, Ccur *C. curvus*, Cfet *C. fetus*, Chom *C. hominis*, Cjej *C. jejuni*, Cmuc *C. mucosalis*, Crec *C. rectus*, Csho *C. showae*, Cspu *C. sputorum*

The phylogenetic trees generated from the 16S rRNA sequences of the 33 genomes are shown in Figure 19. When all 31 genomes were included, the *C. concisus* genomes clustered in one branch. When the related species were removed three main branches were observed. Branch one contained eight 16S rRNA genes from *C. concisus* GS2 genomes (including both 16S rRNA gene constructs from UNSW2) and one 16S rRNA gene construct from the GS1 genome Lasto205.94. The second branch contained the three 16S rRNA genes from *C. concisus* GS2 genomes. Branch three contained 12 16S rRNA genes from *C. concisus* GS1 genomes, including the other construct from Lasto205.94. All of the *C. concisus* percentage identity pairwise results (both intra-GS and inter-GS) exceeded 99% (data not shown).

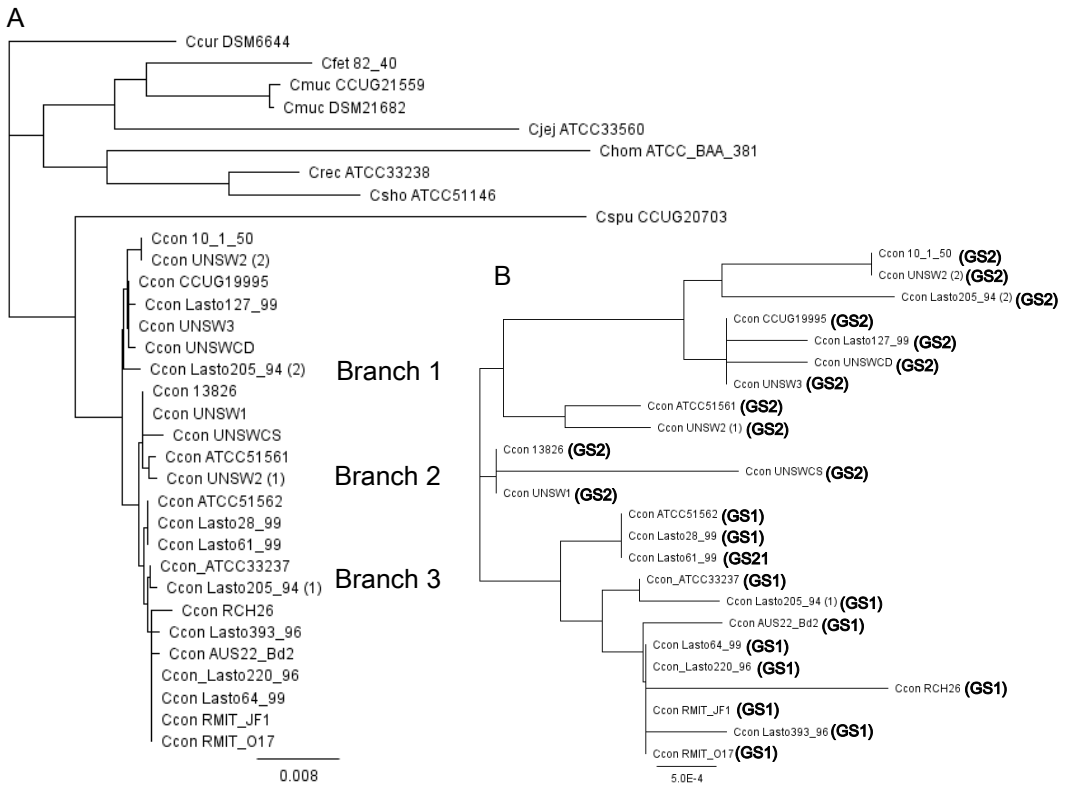


Figure 19: Phylogenetic Trees for *C. concisus* and Related Species based on 33 16S rRNA Gene Sequences

Two *C. concisus* strains had two distinct 16S rRNA sequences.

Panel A. 24 *C. concisus* sequences (representing 22 strains) and 9 related species. Panel B. 24 *C. concisus* sequences (representing 22 strains).

Ccon *C. concisus*, Ccur *C. curvus*, Cfet *C. fetus*, Chom *C. hominis*, Cjej *C. jejuni*, Cmuc *C. mucosalis*, Crec *C. rectus*, Csho *C. showae*, Cspu *C. sputorum*

3.4.3 Identification of Possible Genomospecies-Specific Genes

Fourteen GS1-specific CDS were identified in the *C. concisus* LS-BSR results using the `listAandB.pl` script compared to nine identified using Roary (Page et al. 2015) and reported by Chung et al. (2016) (Table 5). Similarly, 27 GS2-specific CDS were identified using LS-BSR compared to 14 using Roary (Table 6). The specificity of these CDS/genes were evaluated using BLAST searches of three local databases (EpsiloFsa, KrunoFSa and RMITConcisus) and three CDS had Grades of over 96% to non-target CDS. CCON33237_1772, a 159 nt CDS annotated to produce a Hypothetical protein, was identified using Roary as GS1-specific but had Grades of over 96% to GS2 CDSs. CCC13826_0466, a 1041 nt CDS annotated to produce a Luciferase family protein, was identified using LS-BSR and the `listAandB.pl` script as

GS2-specific but had significant similarity to GS1 and *C. showae* CDSs. CCC13826_1636, a 711 nt CDS annotated to produce Aquaporin Z, was identified using Roary as GS2-specific but had significant similarity to *C. showae* CDSs.

Table 5: Comparison of the *C. concisus* GS1-Specific CDS Identified using LS-BSR and Roary

Locus_tag	Gene Products	COG[a]	Identified By	Length (nt)	Minimum Target Grade (%)[b]	Maximum Non-Target Grade (%)[b]
CCON33237_1253	Transcriptional regulator, Crp family	COG0664; K	Both	645	98.5	21.0
CCON33237_1169	Phosphate ABC transporter, periplasmic substrate-binding protein PstS	COG0226; P	Both	873	97.0	23.5
CCON33237_1170	Phosphate ABC transporter, permease protein PstC	COG0573; P	Both	858	96.4	21.5
CCON33237_1171	Phosphate ABC transporter, permease protein PstA	COG0581; P	Both	858	96.6	19.4
CCON33237_1254	Tellurite-resistance/dicarboxylate transporter, TDT family	COG1275; P	Both	960	97.9	58.5
CCON33237_1252	Putative NADH dehydrogenase	COG0446; S	Roary	1338	84.0	77.8
CCON33237_1262	Sel1 domain repeat-containing protein	COG0790; S	LS-BSR	444	94.9	24.6
CCON33237_0317	Hypothetical protein	S	LS-BSR	546	97.9	83.5
CCON33237_1832	Nitrous oxide reductase accessory protein (NosL)	S	LS-BSR	1122	95.6	88.3
CCON33237_0734	Hypothetical protein	T	Roary	747	97.2	87.7
CCON33237_0883	Transporter, AbgT family		Both	1539	97.9	88.7

Locus_tag	Gene Products	COG[a]	Identified By	Length (nt)	Minimum Target Grade (%)[b]	Maximum Non-Target Grade (%)[b]
CCON33237_1674	Hypothetical protein		LS-BSR	495	97.5	85.0
CCON33237_1772	Hypothetical protein		Roary	159	96.8	**96.2**
	Hypothetical protein[c]		LS-BSR	183	94.0	no non-target hits
	Hypothetical protein[c]		LS-BSR	162	98.5	60.5
	Hypothetical protein[c]		LS-BSR	183	94.3	31.5
	Hypothetical protein[c]		LS-BSR	204	85.1	39.3

[a] Clusters of Othologous Groups (COG) numbers (if available) followed by functional groups where K = Transcription; P = Inorganic ion transport and metabolism; S = Function unknown; and T = Signal transduction mechanisms. [b] Grade is a percentage calculated by Geneious by combining the query coverage, e-value and identity with weights of 0.5, 0.25 and 0.25, respectively (Biomatters Ltd 2015) and aides in the identification of the longest, highest identity hits. [c] No gene annotated on GenBank file. Prokka annotation was Hypothetical protein

Table 6: Comparison of the *C. concisus* GS2-Specific CDS Identified using LS-BSR and Roary

Locus_tag	Gene Products	COG[a]	Identified By	Length (nt)	Minimum Target Grade (%)[b]	Maximum Non-Target Grade (%)[b]
CCC13826_0436	Oxidoreductase, FAS/FMN-binding	COG1902; C	Both	1023	96.3	81.9
CCC13826_1636	Aquaporin Z	COG0580; G	Roary	711	85.9	**96.1**
CCC13826_1452	Translation initiation inhibitor	COG0251; J	LS-BSR	384	97.7	83.0
CCC13826_0272	DNA-3-methyladenine glycosylase I	COG2818; L	Roary	558	95.5	29.3
CCC13826_1511	Asparate racemase	COG1794; M	Both	693	97.4	84.4
CCC13826_1584	Twitching motility protein	COG2805; N, U	Both	1161	97.4	76
CCC13826_1830	FlavoCytochrome c heme subunit	P	LS-BSR	438	94.4	89.6
CCC13826_0895	Periplasmic protein	COG3672; S	Both	627	96.8	80.6
CCC13826_1263	Rhomboid family protein	COG0705; S	Roary	576	96.8	44.1
CCC13826_1451	MukF protein	COG2964; S	LS-BSR	621	95.9	21.3
CCC13826_1702	LemA protein	COG1704; S	Both	564	97.7	79.2
CCC13826_2180	Beta-lactamase HcpA (Cysteine-rich 28 kDa protein)	COG0790; S	Both	960	97.2	23.5
CCC13826_1703	Hypothetical protein	S	LS-BSR	891	96.4	26.1
CCC13826_0721	PAS/PAC sensor signal transduction histidine kinase	T	Both	741	97.3	53.1

Locus_tag	Gene Products	COG[a]	Identified By	Length (nt)	Minimum Target Grade (%)[b]	Maximum Non-Target Grade (%)[b]
CCC13826_0177	Na+/H+ antiporter NhaC		Both	1413	96.0	84.3
CCC13826_0178	Beta-aspartyl peptidase		Both	1131	96.5	86.6
CCC13826_0466	Luciferase family protein		LS-BSR	1041	95.0	**96.7**
CCC13826_0565	Hypothetical protein		LS-BSR	516	96.0	85.1
CCC13826_0683	Conserved hypothetical protein		LS-BSR	372	95.7	85.7
CCC13826_0780	Hypothetical protein		LS-BSR	519	96.5	68.2
CCC13826_1402	Glyoxalase II		Both	834	97.6	23.6
CCC13826_1571	Hypothetical protein		LS-BSR	612	95.1	84.3
CCC13826_1698	Hypothetical protein		LS-BSR	540	93.8	24.8
CCC13826_1704	Hypothetical protein		LS-BSR	654	95.6	24.8
CCC13826_1813	PQQ enzyme repeat domain protein		LS-BSR	930	96.9	72.6
CCC13826_1826	Conserved hypothetical protein		LS-BSR	519	94.2	26.9
CCC13826_1877	Conserved hypothetical protein		LS-BSR	540	94.6	85.3
CCC13826_2181	Putative beta-lactamase HcpC (Cysteine-rich protein)		LS-BSR	441	96.5	27.6
CCC13826_2285	Putative lipoprotein		LS-BSR	462	96.6	27.1

[a] Clusters of Othologous Groups (COG) numbers (if available) followed by functional groups where C = Energy production and conversion; G = Carbohydrate transport and metabolism; J = Translation, ribosomal structure and biogenesis; L = Replication, recombination and repair; M = Cell wall/membeane/envelope biogenesis; N = Cell motility; P = Inorganic ion transport and metabolism; S = Function unknown; and U = Intracellular trafficking, secretion, and vesicular transport. [b] Grade is a percentage calculated by Geneious by combining the query coverage, e-value and identity with weights of 0.5, 0.25 and 0.25, respectively (Biomatters Ltd 2015) and aides in the identification of the longest, highest identity hits.

3.5 Discussion

A plethora of tools is now available for evaluating the phylogenetic relatedness of bacterial strains based on WGS. For two of these tools, ANI (with Tetra) and GBDP, species boundary values have been proposed (Auch, Klenk, and Goker 2010, Konstantinidis and Tiedje 2005, Richter and Rossello-Mora 2009, Rossello-Mora and Amann 2015). All of the *C. concisus* intra-GS ANIb and ANIm results were between 93% and 96%, the "fuzzy zone" where species circumscription cannot be assured (Rossello-Mora and Amann 2015). Conversely, all of the *C. concisus* inter-GS ANIb and ANIm results fall well below this zone, with results similar to the *C. rectus/C. showae* inter-species pair, suggesting the two groups are from distinct genomospecies. It has been proposed that Tetra can complement ANI in the assignation of species identification such that ANI >96% and Tetra >0.999 is considered to provide assurances of species circumscription that can be verified by phenotypic studies (Rossello-Mora and Amann 2015). None of the *C. concisus* pairs fulfil this criteria since none of the ANI results exceeded 96%. However, 70% of the intra-GS pairs did exceed 0.999 for Tetra.

Richter and Rossello-Mora (2009) earlier proposed a species boundary of ANI >95-96% and Tetra >0.99. Almost all (98.9%) of the *C. concisus* GS2, and a small proportion (4.9%) of the GS1 intra-GS results exceeded this species boundary when the ANIm algorithm was used, supporting at least two distinct genomospecies within the 22 *C. concisus* genomes.

GDBP aims to provide a result equivalent to DNA-DNA reassociation so has a proposed species boundary of 70% (Auch, Klenk, and Goker 2010). Only the *C. mucosalis* intra-species results exceed this value using the recommended formula 2. For formulas 1 and 3, all of the GS1 and 15.6% of the GS2 intra-GS GBDP pairwise results exceeded 70% but 5.8% of the *C. concisus* inter-GS results also exceeded 70% using formula 1. For this reason, formula 3 was preferred for this dataset. This was also the optimal formula when GBDP was performed on selected complete genomes of well-characterised members of *Campylobacteraceae* and *Helicobacteraceae* (On et al. 2017). All of the heatmaps for GBDP clustered the same *C.*

concisus genomes into GS1 and GS2, as observed for ANI and Tetra, with the related species in a separate branch. This adds support to at least two genomospecies within *C. concisus*, even if the proposed species boundary isn't consistently achieved.

FFP separated the 22 *C. concisus* genomes into two clear clusters with the genomes representing related species forming another cluster. The members of the two *C. concisus* clusters are the same as observed for ANI, Tetra and GBDP. This alignment-free technique has been used to evaluate the phylogenetic relationship of seven gastric *Helicobacter* species (van Vliet and Kusters 2015).

rMLST, OrthoMCL and COGs analyses all separated the 22 *C. concisus* genomes into the same two genomospecies generated by ANI, Tetra, GBDP and FFP. rMLST has been proposed to provide phylogenetic information over a very broad taxonomic range (Jolley et al. 2012) and has been shown to provide species-level identification as well as sub-species clustering for the closely related species *C. jejuni* and *C. coli* (Read et al. 2013) and providing data towards the classification of a new *Campylobacter* species (Grange et al. 2016). rMLST has also been applied to the classification of new species within the Enterobacteriales order (Facey et al. 2015) and provided clarification of classification within the *Neisseria* genus (Bennett et al. 2012, Bennett, Jolley, and Maiden 2013). OrthoMCL has been used to evaluate the pan-genome of the *Bacteroides* genus (Yoshizaki et al. 2014) and the distribution of COGs in the core- and pan-genomes of four genera has also been compared (Lukjancenko, Ussery, and Wassenaar 2012) although the comparisons did not extend to evaluating the phylogenetic relationship of species within each genus.

FFP and rMLST do not have numerical cut-offs for species circumscription but rely on clustering the genomes under investigation to provide evidence of phylogenetic relatedness. OrthoMCL and COGs were used as the subjects of an exploratory data analysis to evaluate whether the separation of the GS was maintained when a functional analysis was performed. The NeighborNets generated from the analysis of the 22 *C. concisus* genomes using these four approaches to phylogenetic inference resulted in the same *C. concisus* genomospecies being

observed as those seen in the ANI, AF, Tetra and GBDP with the type strain ATCC 33237^T representing GS1 and CCUG 19995 representing GS2.

The GC content range calculated directly from the sequenced *C. concisus* genomes (37.36-39.81%) are well within the GC content range (37-41 mol%) estimated from the physical properties induced in extracted or digested genomic DNA (Meier-Kolthoff, Klenk, and Goker 2014) and reported as molar percentages (mol%) (Vandamme et al. 2005b). Meier-Kolthoff, Klenk, and Goker (2014) have proposed that GC differences should be no more than 1% within species if using genome sequences. The median GC contents of the two *C. concisus* genomospecies differ by 2% and all of the intra-GS pairwise GC differences were <1 while all of the inter-GS pairwise GC differences were >1%. The GC content calculated directly from the ATCC 33237^T (37.62%) and CCUG 19995 (39.39%) genomes were similar to the results observed using thermal denaturation (37.9 mol% and 39.9 mol%, respectively) (Vandamme et al. 1989) which also differ by 2%. The pairwise GC differences between the two *C. mucosalis* genomes and some of the *C. concisus* GS1 genomes were also <1%. This is a reflection of the GC contents of *C. mucosalis* genomes which were 36.64% and 36.54% for DSM 21682^T and CCUG 21559, respectively. The type strain result calculated directly from the genome is similar to the 37.1 mol% calculated by thermal denaturation (Vandamme et al. 1989). These results are within the range reported for *C. mucosalis* (36-38 mol%, (Vandamme et al. 2005b)). The similarity of GC content results between *C. concisus* GS1 and *C. mucosalis* demonstrates that genomes from different species can have similar GC content so this analysis can not be used in isolation to provide species identification. The relationship between genomospecies and both genome size and N_{50} were also evaluated using the 22 *C. concisus* genomes but there were no differences observed between the genomospecies for either of these measures (data not shown).

Larsen et al. (2014) compared the ability of SpeciesFinder, rMLST, TaxonomyFinder and KmerFinder to predict on assembled draft or complete genomes from 695 isolates representing 81 genera and 149 species. They found that SpeciesFinder, which uses full length 16S rRNA

sequence, and rMLST performed poorly (Larsen et al. 2014). rMLST performed somewhat better for 10,407 draft genomes generated from the Sequence Read Archive but the method consistently made incorrect identification for a number of closely related species for both groups of genomes (Larsen et al. 2014). Since Larsen and colleagues generated their own rMLST database, the poor performance of this method in this comparison may relate to the criteria applied by the authors. Genomes can now be uploaded to the PubMLST rMLST website[44] where a predicted taxon is provided and a list of the best match for each gene is generated. Interestingly, KmerFinder, which uses a similar principle to FFP but using a kmer (*k*-mer) length of 16 and only a proportion of the generated kmers, performed the best (Larsen et al. 2014). Both rMLST and FFP separated the 31 genomes in this study into three main branches relating to the two *C. concisus* genomospecies and the related species which was concordant with the ANI, Tetra, GBDP, OrthoMCL, and COGs results.

The phylogenetic tree generated using 16S rRNA gene sequences extracted from all 31 genomes demonstrated that there was little diversity within *C. concisus* 16S rRNA genes. Even when the 16S rRNA sequences from related species were excluded, the separation of the *C. concisus* genomospecies observed for the genome-wide analyses was poor. The observation of fragmented 16S rRNA in two of the *C. concisus* genomes may have affected the resolution of the phylogenetic trees generated from these genes. Amongst the eight *C. concisus* genomes generated in this project, the only CDS annotated across a contig boundary was the 16S rRNA gene of Lasto205.94 suggesting this is not a common occurrence. This could, however, be further improved by substituting the 200 Ns used to join the contigs of draft genomes with the spacer sequence described by Altermann, Lu, and McCulloch (2017) which includes stop codons across all six reading frames.

All of the *C. concisus* 16S rRNA pairwise percentage identity results exceeded 99% demonstrating that this gene has poor discriminatory power for these genomes. Chung et al.

[44] https://pubmlst.org/rmlst/

(2016) also observed little diversity in the 16S rRNA gene sequence of *C. concisus* isolates. The discriminatory power of the 16S rRNA gene sequence has previously been shown to be poor for some species within the *Campylobacter* genus (Gorkiewicz et al. 2003, Hansson et al. 2008, Korczak et al. 2006). In addition, WGS-based analyses such as ANI are considered superior to 16S rRNA sequence analysis for studying phylogeny because they are based on a much larger part of the genome and have better resolution for discriminating both distantly and closely related bacteria (Vandamme and Peeters 2014). Complete genomes provide the best starting information for WGS-based analyses but as long as finishing genomes is out of the reach of most laboratories, draft genomes provide an adequate starting point for these analyses.

The proposed species boundaries for ANI, Tetra, GBDP were not consistently exceeded for *C. concisus* intra-group pairs however two clusters containing the same strains were consistently produced by all methods except 16S rRNA gene sequencing. Previous studies into the genetic diversity of *C. concisus* have observed significant heterogeneity, even below the level of genomospecies. Examples of this diversity include 51 PFGE patterns from 53 isolates (Matsheka et al. 2002), 62 AFLP types from 62 isolates (Aabenhus et al. 2005), 21 AFLP types from 22 isolates (Kalischuk and Inglis 2011), 47 AFLP types from 47 isolates (On, Siemer, et al. 2013), 26 ST from 70 isolates (Ismail et al. 2012) and 30 ST from 37 isolates (Chung et al. 2016) using a six-gene MLST scheme, 66 ST from 70 isolates using a seven-gene MLST scheme (Miller et al. 2012), and 34 ribotyping patterns and 37 random amplified polymorphic DNA patterns from 43 isolates (Engberg et al. 2005). A 23S rRNA PCR method originally described using one forward and two reverse primers for the detection of *C. concisus* (Bastyns et al. 1995) and later modified by the conversion to two single-reverse primer PCR reactions separated most *C. concisus* isolates into two genomospecies (Engberg et al. 2005, Istivan et al. 2004, Kalischuk and Inglis 2011, On, Siemer, et al. 2013). Alignment to these reverse primers was also used in section 4.4.1 Pan-genomic Analysis to assign *C. concisus* genomes to temporary genomospecies. The temporary genomospecies assignments matched the groups

generated by ANI, Tetra, GBDP, FFP, rMLST, OrthoMCL and COGs. The separation of genomospecies observed for the Roary core-genome phylogenetic tree was also congruent with the phylogenetic tree generated using 23S rRNA gene sequences of 37 *C. concisus* isolates (Chung et al. 2016). Where isolates are shared between published studies, and including the work presented in this thesis, the genomospecies assignments were congruent.

It has been proposed that where strains exhibiting 16S rRNA gene sequence identities of over 97%, DNA reassociation studies should be undertaken as the gold standard for species delineation (Stackebrandt and Goebel 1994). The degree of binding between the pairs ATCC 33237^T (Group 1)/CCUG 19995 (Group 2) and ATCC 51561 (CCUG 20034)/CCUG 19995 (both Group 2) have been reported as 46% and 77%, respectively (Vandamme et al. 1989) providing support of the existence of two species within *C. concisus*. Given that a phenotypic difference between the two *C. concisus* groups has not yet been identified, the two groups are considered genomospecies.

Six of the 12 genomes in GS1 were from strains previously assigned to GS1 by AFLP, one (Lasto393.96) was assigned to GS6 and five (ATCC 51562, AUS22-Bd2, RCH26, RMIT-JF1 and RMIT-O17) has not been previously analysed using AFLP (On, Siemer, et al. 2013). Of the three GS2 genomes that have previously been assigned to genomospecies using AFLP, two were GS2 and one (Lasto127.99) was GS5 (On, Siemer, et al. 2013). The Lasto127.99 (AFLP-GS5) and Lasto393.96 (AFLP-GS6) AFLP results were singletons with atypical banding patterns. Given that AFLP is a semi-random sampling of the genome it may be expected that aberrant results are sometimes observed. Interestingly, Lasto393.96 (GS1 and AFLP-GS6) was positive for the MUC+CON2 Bastyns 23S rRNA PCR (On, Siemer, et al. 2013) but showed higher similarity to the CON1 primer and generated the MLPA products expected from *C. concisus* GS1 in Figure 20. The AFLP and DNA reassociation results demonstrate general agreement between *in silico* and wet-laboratory comparisons supporting the existence of at least two *C. concisus* genomospecies and reinforcing the groups identified in this study. Two *C. concisus* genomospecies are generally recognised using multiple gene sequencing methods

(Chung et al. 2016, Huq et al. 2017, Miller et al. 2012, Wang et al. 2017), however there is some evidence that additional genomospecies may exist for this species (Mahendran et al. 2015, Nielsen, Nielsen, and Torpdahl 2016).

The genomospecies-specific CDS generated by LS-BSR (with `listAandB.pl`) and Roary overlap. When the minimum BSR for target genomes was increased from 0.8 to 0.9, the number of GS2-specific CDS identified by LS-BSR (with `listAandB.pl`) was reduced from 27 to 23 without any loss of CDS also identified using Roary. Although a similar adjustment with respect to GS1 genomes did not result in any decrease in the GS1-specific CDS identified by LS-BSR (with `listAandB.pl`), it is possible that the decrease observed for GS2-specific CDSs reflects a difference in the algorithms that contributes to the differences in the CDS identified as specific for a particular taxon. The variation in genomes making up the two pan-genomic studies is also likely to have affected the CDS identified. Nine genomes (2 GS1 and 7 GS2) were common to both studies, eight genomes were unique to LS-BSR and 27 were unique to Roary.

Four of the GS1-specific CDS only identified using LS-BSR (with `listAandB.pl`) were not annotated in the GenBank files but were annotated as hypothetical proteins by Prokka. All of these CDS were <220 nt. One of the GS1-specific CDS only identified using Roary was also <220 nt. The locus_tag for this CDS is not included in the version of the ATCC 33237^{T} complete genome available from GenBank on 14th May 2017 but was included in an annotation spreadsheet shared by Dr William Miller (Miller 2017). It is therefore possible that the CDS used as a query in the BLAST searches was not the same CDS identified by Roary. Short CDS are less likely to encode a protein. If CDSs less than 500 nt long were excluded to increase the confidence that only protein-encoding CDS are included, 10 GS1-specific CDS were identified; six were identified by both methods, two only by LS-BSR (with `listAandB.pl`) and two only with Roary. Similarly, 22 GS2-specific CDS had BSR results of at least 0.9 for all GS2 genomes and were at least 500 nt long with 11 being identified by both methods, eight only by LS-BSR (with `listAandB.pl`) and three being identified only by Roary. Three CDSs identified as GS-specific using Roary or LS-BSR (with `listAandB.pl`) were found to have significant sequence

similarity (Grades of ≥96.0%) to CDSs of non-target genomes when they were used as queries for BLAST searches of databases containing Epsilonproteobacteria genomes that were not included in the original analyses. These results demonstrate that the identification of taxon-specific CDSs is dependent on the dataset used to generate them and that the specificity of CDSs should be challenged using both taxonomically broader databases and as new data becomes available.

Within the two genomospecies there are no clear clusters of genomes that are observed across multiple analyses suggesting that any groupings below this level are not phylogenetically meaningful. The draft GenBank genome 10_1_50, which has only been identified to the genus level and tentatively assigned to *C. concisus* on the basis of the sequence similarity with the *C. concisus* GS2 23S rRNA sequence (Figure 22) and taxon-specific genes, clustered within *C. concisus* GS2 for all analyses. This demonstrates that whole genome sequencing may be a cost effective method for tentatively assigning partially characterised isolates to a taxonomic group, especially for laboratories not specialising in the phenotypic identification of the genus of an isolate of interest. They also provide useful phylogenetic information that can help inform classification.

3.6 Contributions

Angela Cornelius performed all of the ANI, Tetra, GBDP, FFP, LS-BSR, QUAST and the analyses undertaken in and around Geneious. She designed the `listAandB.pl` script written by Associate Professor Biggs and generated the Prokka annotations, heat maps, NeighborNets, boxplots and 16S rRNA sequence alignments. Angela also interpreted all of the data and generated all of the tables and figues in this chapter.

Associate Professor Patrick Biggs assembled the eight *C. concisus* whole genome sequences, wrote the Perl script used to join the genomes and generate the Prokka annotations, and conducted the rMLST, OrthoMCL and COGs analyses.

Dr Darren Smalley prepared the cultures and performed the DNA extractions for the *C. concisus* strains for which whole genome sequences were generated.

The whole genome sequencing of the eight *C. concisus* strains was performed by New Zealand Genomics Limited (NZGL).

Chapter 4: Multiplex Ligation-dependent Probe Amplification

4.1 Abstract

The Epsilonproteobacteria class includes three genera containing taxa known or suspected of causing human gastroenteritis. Although it is possible to detect these taxa using molecular techniques such as PCR, there is currently no single assay for the detection of taxa across the three genera. Large scale-BLAST score ratio (LS-BSR) analysis was used to evaluate the pan-genomes of the three genera *Arcobacter, Campylobacter* and *Helicobacter*. Candidate coding sequences (CDS) for MLPA design were identified for 23 taxa on the basis of intra- and inter-taxon similarity, length and putative function. Delta-bitscore (DBS), a profile-based homology scoring method, was used to identify candidate genes for one taxon and published PCR primers were used to identify suitable genes for four taxa. Multiplex ligation-dependent probe amplification (MLPA) probes were designed and manufactured for these 28 taxa and tested against a collection of 127 DNA extracts representing 62 Epsilonproteobacterial taxa and 15 human enteric bacterial pathogens. The MLPA products were detected using capillary electrophoresis and analysed in BioNumerics using the AFLP module. Band matching assigned peaks to the probes on the basis of length and the results were exported as common-separated values. Concordant results were obtained for the majority of probes and DNA extracts. However, two probes (Clariconcheus and Hpullorum) failed to detect their target taxa and three other probes (ClariUPTC, Cupsailensis and Csubantarcticus) need to be replaced or modified to ensure the products detected are concordant with the taxa present in a sample. In addition, the MLPA assay appears to be sensitive to EDTA and optimisation of band matching settings was required to minimise the assigning of “aberrant” peaks from the true positive products being read as the next smaller product. Although not all of the probes provided sensitive, specific and repeatable results, and further optimisation is required, the MLPA assay

shows promise for the identification, and simultaneous detection, of a range of Epsilonproteobacteria taxa.

4.2 Introduction

Identification is the process of assigning an unknown item to a unit that has previously been classified and named (Cowan 1965). A bacterial isolate is considered identified only if an acceptable level of matching between the isolate and a known taxon is achieved (On 2005).

Traditionally, taxa within the four animal-associated Epsilonproteobacteria genera *Arcobacter, Campylobacter, Helicobacter* and *Wolinella* have been identified using standard phenotypic characteristics involving mostly physical characteristics or biochemical tests (On et al. 2005, Vandamme et al. 2005b, c) but it has been noted that this process can be problematic for this bacterial class due to their relatively inert nature and complex taxonomy (On 2013).

Amplification and detection of taxon-specific DNA segments using the polymerase chain reaction (PCR) has emerged as a popular method for the molecular identification of *Arcobacter* (Collado and Figueras 2011)*, Campylobacter* (On 2013) and *Helicobacter* (On et al. 2005) species. These PCR methods are relatively quick and easy to perform but can generally only identify a small number of taxa.

Identification methods require pure isolates but the varied and complex nutrient and atmospheric requirements, coupled with the often slow growth rates, of Epsilonproteobacteria (Kroger et al. 2005, On et al. 2005, Robertson et al. 2005, Stolz et al. 2005, Vandamme et al. 2005b, c) make isolation and identification of species within this class challenging.

Molecular methods offer the potential to detect a broad range of bacterial species in a manner that is not biased by culture conditions. A variety of methods have been published for the molecular detection of the established or potential human pathogenic Epsilonproteobacterial species in human stool samples. These methods include loop-mediated isothermal amplification (Pham et al. 2015, Ushijima et al. 2014, Yamazaki et al. 2008, Yari et al. 2016), PCR (Abdelbaqi et al. 2007, Collado et al. 2013, Oyama et al. 2012, Samie et al. 2007) and PCR-denaturing

gradient gel electrophoresis (PCR-DGGE, (Cornelius et al. 2012, Vandenberg et al. 2013)) assays and generally detect either a single taxon, or less than 10 taxa per reaction. The PCR-DGGE method, although useful for identifying multiple Epsilonproteobacterial taxa in a single assay, is time-consuming and not suitable for routine use. Molecular methods for the detection of a range of gastrointestinal pathogens have been reported including commercially available multiplex PCR (Anderson, Buchan, and Ledeboer 2014, Koziel et al. 2013, Zhang, Morrison, and Tang 2015), TaqMan Array Card (Liu et al. 2014) and DNA microarray assays (Donatin et al. 2013, You et al. 2008). These assays, although broad in application, detect a maximum of five Epsilonproteobacterial species or provide only genus-level detection of one Epsilonproteobacterial genus (*Campylobacter*).

Multiplex ligation-dependent probe amplification (MLPA) is a modification of PCR (Figure 6) that allows up to 40 genes to be targeted within a single reaction (Schouten et al. 2002). Each MLPA probe consists of two oligonucleotides, the left probe oligonucleotide (LPO) and right probe oligonucleotide (RPO), that are ligated to each other when both the left hybridization sequence (LHS) and right hybridization sequence (RHS) are hybridised to a target sequence (Schouten et al. 2002). MLPA probes are generally longer than PCR primers and the assay uses a relatively high annealing temperature to assist with specificity. The 3' end of the LHS is the most sensitive to mismatches and this has been exploited for single nucleotide polymorphism (SNP) analysis (Bergval et al. 2012). All MLPA probes have the same PCR primer sequence X at the 5' end of the LPO and PCR primer sequence Y at the 3' end of the RPO permitting simultaneous PCR amplification with a single primer pair (Schouten et al. 2002). The MLPA probes are designed to have unique amplification product lengths allowing recognition of each target sequence on the basis of the size of the amplification products (Schouten et al. 2002). This is achieved by the inserting defined lengths of DNA from the bacteriophage M13 between the RHS and PCR primer sequence Y on the RPO. It is also possible to make MLPA probes with three probe segments. The third probe segment, called a spanning probe (SP), anneals between the LPO and RPO. The requirement for two ligation

events increases the specificity of three-segment probes. MLPA assays can be designed to be undertaken in a single working day and use relatively basic molecular biology equipment (Cornelius et al. 2014).

Unbiased identification of possible gene targets for MLPA requires the comparison of genomes from all target taxa. An ever increasing number of whole genome sequences, in draft and complete form, are publicly available from the International Nucleotide Sequence Database Collaboration (INSDC[45]) and online data repositories such as DRYAD[46]. INSDC is a collaboration between the National Center for Biotechnology Information's GenBank[47], the DNA DataBank of Japan (DDBJ[48]) and the European Nucleotide Archive[49] within the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI[50]) that provides free and unrestricted public access to nucleotide sequence and annotation information submitted to all three databases. These sequences can be downloaded for use by individual researchers.

Within any given taxon there is a core genome, which consists of genes shared by all strains belonging to this taxon, and a dispensable or accessory genome, which includes genes that are present in some, but not all, strains (Tettelin et al. 2005). The sum of the core and accessory genomes is called the pan-genome (Figure 7) (Tettelin et al. 2005). Ideal candidate genes for MLPA probe design are core to the target taxon and absent from the pan-genomes of all non-target taxa.

This chapter uses the pan-genome of the Epsilonproteobacteria class to identify genes, or gene variants, that are conserved in a chosen taxon, and absent from all other taxa. MLPA probes were designed to specifically detect these genes, or gene variants, and then the resulting assay was evaluated using a collection of control DNA representing the three

[45] www.insdc.org

[46] https://datadryad.org

[47] www.ncbi.nlm.nih.gov/genbank

[48] http://www.ddbj.nig.ac.jp/

[49] http://www.ebi.ac.uk/ena

[50] http://www.ebi.ac.uk/

Epsilonproteobacterial genera that contain human pathogens as well as a small collection of other bacterial human enteric pathogens.

4.3 Materials and Methods

4.3.1 Epsilonproteobacterial Genomes

A total of 799 Epsilonproteobacterial genomes, 150 of which were complete and 649 were in draft form, were publicly available from GenBank on 24th December 2014. These genomes, including any associated plasmids, were downloaded into Geneious[51] R6.1.7 and FASTA files for each genome were exported. Four additional collections of genomes for taxa from this bacterial class were similarly saved in Geneious R6.1.7 and FASTA files exported. The first collection contained 80 draft *C. jejuni* and *C. coli* genomes as reported by Sheppard, Didelot, Meric, Torralbo, et al. (2013) and available via DRYAD (Sheppard, Didelot, Meric, Torralba, et al. 2013). The second collection contained 43 complete genomes from *Arcobacter* and *Campylobacter* strains kindly supplied by Dr William Miller (USDA) and the third contained 9 draft genomes representing three potentially new *Campylobacter* species and was kindly supplied by Dr Patrick Biggs and colleagues (Massey University). The eight draft *C. concisus* genomes generated in 3.3.1 Whole Genome Sequences made up the fourth collection of genomes. All of the sequences associated with each genome were concatenated, joined and the resulting 939 genomes were annotated using Prokka (version 1.8 (Seemann 2014)) to provide consistent annotations. The Prokka-generated GenBank (.gbk) files were imported into Geneious and BLAST databases generated for the nucleotide (EpsiloFsa) and translated amino acid (EpsiloFaa) sequences of all 939 genomes. A group of three *C. helveticus* and three *C. upsaliensis* draft genomes was subsequently supplied by Dr Krunoslav Bojanic (Massey University). These genomes were processed as above and a separate nucleotide BLAST database (KrunoFsa) generated. Likewise, a set of four draft *C. concisus* genomes was

[51] http://www.geneious.com/

subsequently supplied by Dr Mohsina Huq (RMIT University). These genomes were processed as above and a separate BLAST database (RMITConcisus) generated. In most cases, the probe design preceded the availability of additional genomes so two new BLAST databases were generated to provide supplementing information. The complete list of genomes and the BLAST databases they were included in is available in Supplementary File 2.

4.3.2 Pan-genomic Analysis and Target Gene Selection

A list of target taxa for MLPA development was generated that included species and subspecies with an established, or suspected, role in human gastroenteritis. *Arcobacter,* *Campylobacter* and *Helicobacter*, the three Epsilonproteobacterial genera representing the taxa of interest for this assay, were also included in the MLPA development list. Taxa phylogenetically-related to the important human pathogen *C. jejuni* were identified from the 16S rRNA gene tree reported by On and Cornelius (2016) and included in the list of taxa for MLPA development.

Large Scale BLAST Score Ratio (LS-BSR, (Sahl et al. 2014)) was used to identify predicted coding sequences (CDS) that were core for each taxa in the following six analyses: i) the entire set of 939 Epsilonproteobacterial genomes, ii) the 21 *Arcobacter* genomes, iii) 329 *Campylobacter* genomes, iv) 428 *Helicobacter* genomes, v) 17 *C. concisus* genomes, and vi) 6 *C. lari*-group genomes. BSR is calculated by dividing the BLAST score for a query peptide by the self-BLAST score of the reference peptide (Rasko, Myers, and Ravel 2005). This normalises all results to the range 0 to 1 where a score or 1 is a perfect match for the length of the reference peptide and a score of 0 indicates no match between the query and reference peptides (Rasko, Myers, and Ravel 2005). This allows comparisons of peptide similarity using universal cutoffs (Rasko, Myers, and Ravel 2005). All six LS-BSR analyses used nucleotide sequences as the input data and the default BSR cut-off of 0.9.

Genomes from species not validly described on 24th December 2014 as well as genomes from strains only identified to the genus level were ignored when the LS-BSR results were

interrogated. Genomes from strains modified in the laboratory (including those passaged through laboratory animals, for example) were also ignored. To further enhance the quality of the data evaluated, duplicate genomes for the same strain were ignored. In this case preference was given to complete, published, GenBank genomes. Where multiple draft genomes were available for the same strain, preference was generally given to those with fewer contigs.

Seventeen (17) *C. concisus* genomes were available for the *C. concisus* LS-BSR analysis but only 10 had genomospecies previously established using amplified fragment length polymorphism (AFLP) (On, Siemer, et al. 2013). In order to approximate AFLP without the benefit of extracted DNA, a Geneious (R6.1.7) alignment with the CON1 and CON2 reverse primers from the *C. concisus* genomospecies-specific PCR (On, Siemer, et al. 2013) and the 23S rRNA genes extracted from the *C. concisus* 17 genomes was generated. The genomes were separated into two groups based on the CON primer for which the 23S rRNA sequence had the highest similarity.

The `compare_bsr.py` script available with LS-BSR was used as a first screen to identify predicted coding sequences (CDS, called centroids in LS-BSR) present in a target set of genomes and absent from the non-target set. When no taxon-specific candidate CDS were identified by `compare_bsr.py`, the matrices generated in LS-BSR were evaluated in R (v 3.2.5, (R Core Team 2016)) using a custom Perl script (`listAandB.pl`), written by Dr Patrick Biggs, to identify CDS of interest. Lists of target (ListA) and non-target (ListB) genomes were generated from the shortlisted genomes for each of the taxa identified for possible MLPA development (see Supplementary File 2). A range of BSR values were evaluated for their ability to generate a modest (<100 members) list of possible target CDS. For the identification of CDS that were core to a genus, a BSR of at least 0.8 in all validly identified taxa within the genus was initially used. BSR values of at least 0.9 and 0.95 were also used where the list of possible CDS exceeded 100. When screening for taxon-specific CDS at the species or subspecies level, a BSR of at least 0.8 in all genomes in the target taxon and 0.4 or less in all genomes from non-target taxa was initially applied. These default BSR values were established using the

Arcobacter LS-BSR data which was small enough for the effect of different values to be evaluated in Excel. More stringent criteria were used for 10 species or subspecies where over 100 CDS were identified using the initial criteria. The results were imported into Excel (2013; Microsoft, Redwood, WA) for collating and additional analyses.

A custom Perl script (`centroidRetrieve.pl`), written by Dr Patrick Biggs, was used to find and extract the nucleotide or translated amino acid sequences for possible target taxa-specific CDSs. These CDSs were imported into Geneious (R8.1.6 or R9.1.7) and a subset were then used as query sequences for BLAST searches of the local and NCBI non-redundant (nr) databases within Geneious (R8.1.6 or R9.1.7). The searches of the local Epsilonproteobacteria-based databases allowed the location of the CDS within the genomes to be identified, as this was not included in the LS-BSR output, it provided an evaluation of the sequence variability within the target taxon and established that the CDS was present in only the target taxon. The genomes ignored during analysis of the LS-BSR data were included in the BLAST databases to evaluate both the quality of these genomes and whether the specificity of the identified CDS was maintained when lower quality and incompletely identified genomes were included. The nr database searches were used to confirm that the CDS had no, or only low sequence similarity with bacteria outside of the Epsilonproteobacteria class.

A profile-based homology scoring method called delta-bitscore (DBS) (Wheeler et al. 2016) was used to identify orthologous proteins that have deviated more from predicted sequence constraints in one *C. concisus* genomospecies (GS) compared to the other. This amino acid based method identified possible genes that could be used to identify *C. concisus* GS1.

The amino acid sequences of CDS annotated by Prokka as hypothetical proteins were used as queries to the Pfam database (Finn et al. 2014) to identify similarity to genes or motifs in common with known proteins.

One shortlisted gene was chosen for MLPA design with priority being given to genes with greater than 500 nucleotides, a putative function assigned by Prokka or Pfam as well as high

intra-taxon and low inter-taxa similarity observed in the BLAST searches. Preference was given to CDS greater than 500 nucleotides as this provides increased confidence that it encodes a protein. Similarily, preference was given to CDS with a putative function annotated as this provides more confidence that the CDS has a biological function. If MLPA design was too challenging for the first gene, an alternative gene was chosen for MLPA design.

Where suitable CDS were not identified for target taxa using `compare_bsr.py` and `listAandB.pl`, the scientific literature was reviewed for genes used for PCR-based detection of the target taxa. The PCR primers targeting *C. coli* and *C. jejuni* from a recent paper that evaluated the specificity and sensitivity of 31 PCR assays targeting these, and related, taxa (On, Brandt, et al. 2013), were used as queries for BLASTn (word size 7) searches of the local EpsiloFsa database in Geneious (R8.1.6 or R9.1.7) to identify *C. coli*- and *C. jejuni*-specific genes suitable for MLPA design.

4.3.3 Design and Manufacture of MLPA Probes

Clustal W/Geneious alignments, including all target versions of the gene as well as any non-target gene sequences showing significant nucleotide similarity, were prepared for each target gene in Geneious (R8.1.6). LHS and RHS were designed using these alignments and the criteria suggested in the Designing synthetic MLPA probes protocol (MRC-Holland 2014) using the criteria summarised in Table 7. The melting temperatures (T_m) of the LHS and RHS were estimated using RawProbe[52] and the minimum free energy (ΔG) of folding for the basic LPO and RPO, which included the LHS/RHS and PCR primer sequences X/Y, were estimated using mfold[53] (Zuker 2003) using 60°C for the folding temperature and 0.35 mM Na^+ concentration.

[52] Raw probe and Designing synthetic MLPA probes (latest version) are available from: http://mlpa.com/WebForms/WebFormMain.aspx?Tag=_zjCZBtdOUyAt3KF3EwRZhAPz9QEm7akikAm7AOEGw1vtZvffaZPOiSig8uqel7Yd

[53] Available at http://unafold.rna.albany.edu/?q=mfold/DNA-Folding-Form

The “Search for motif” feature in Geneious (R8.1.6) was used to identify restriction sites used in manufacturing process within the hybridising sequence (LHS and RHS).

Table 7: Criteria for the Design of Multiplex Ligation-dependent Probe Amplification Probes

Attribute	Criteria	Description
Length of LHS	> 20 nt	Hybridising sequence must be at least 21 nt
Length of RHS	> 20 nt	Hybridising sequence must be at least 21 nt
GC%	50 ± 20	Optimum of 50%, acceptable range 30-70%
T_m (°C)	≥ 68°C	Minimum melting temperature varies by length:
		< 26 nt ≥ 72.7°C (preferably ≥ 74°C)
		26-30 nt ≥ 71°C (preferably ≥ 72.5°C)
		31-35 nt ≥ 71°C
		36-40 nt ≥ 70°C
		41-55 nt ≥ 68°C
First nt of LHS	Not A	Affects strength of signal; CC strongest (use for long probes), T good, avoid A. N.B. could be stuffer
First 4 nt of LHS	< 4 G/C	Maximum of 3 G/C nest to MLPA Target (5’ end of LHS, but could be stuffer)
Last 5 nt of LHS	< 3 G/C	Maximum of 2 G/C in last 5 nt
Last 4 nt of RHS	< 4 G/C	Maximum of 3 G/C nest to MLPA Target (3’ end of RHS, but could be stuffer)
ΔG of LHS	> 0	From mfold, 60°C, 0.35 [Na^+], indicates likely hairpin formation
ΔG of RHS	> 0	From mfold, 60°C, 0.35 [Na^+], indicates likely hairpin formation

Hybridising sequences (LHS and RHS) that fulfilled the criteria above were used as queries for searches of the local EpsiloFsa and NCBI nr databases within Geneious (R8.1.6) using BLASTn and a word size of 7. Hybridising sequences with significant similarity to non-target genes, especially over the ligation site, were discarded. The ligation site is the 3’ end of the LHS and the 5’ end of the RHS.

To facilitate the application of the Epsilonproteobacteria MLPA assay to laboratories without access to capillary sequencing equipment, the ideal MLPA products differed in length by at least 10 nt for products below 200 nt, by at least 12 nt for products between 201 and 310 nt, at least 14 nt for products between 311 and 400 nt, and at least 16 for products over 400 nt. The final LPO and RPO lengths were calculated in Excel by considering these ideal product

lengths, the size of the LHS and RHS sequences and the size of M13 vectors available at MRC-Holland. Up to four nucleotides of non-annealing “stuffer” nucleotides were included to achieve product lengths close to the ideal lengths. Due to the limited number of sizes of vectors available to extend the RPO, some length differences are smaller than planned. The design worksheets and an Excel spreadsheet summarizing all of the MLPA probes in the assay were sent to MRC-Holland for checking. The lengths of a small number of probes were altered in response to suggested improvements from MRC-Holland. The summary spreadsheet included the nucleotide sequences for the LHSs and RHSs along with the lengths for the stuffers, vectors, LPOs, RPOs and MLPA products.

The MLPA probes were manufactured as described by Schouten et al. (2002) using the summary spreadsheet. Briefly, the LPOs and a small number of RPOs were made synthetically and the RPOs were cloned into M13-derived vectors to produce longer probes than can easily be made synthetically. This process takes several months as being a biological process it has a limited throughput and the probes for this project needed to be incorporated into the workflow of MRC-Holland.

Two probemixes (A098 Epsilo A and A099 Epsilo B) were generated by MRC-Holland, each containing probes targeting 14 taxa and 4 DNA quantity fragments (Q-fragments). The probemixes were designed to be used together when capillary electrophoresis would be used for detection of the products. Alternatively, the probemixes can be used separately when less discriminative detection methods such as microfluid electrophoresis systems were being employed. The Q-fragments have lengths of 64, 70, 76 and 82 nt and are included at a low concentration. They are generally not amplified when target sequences are detected but should be visible when no target probes have been amplified. Inhibition of the reaction is suspected when there are no target probes detected and no Q-fragments are visible. Multiple probes, with slight variations in the nucleotide sequences, were included to account for genetic variation observed in the target region for some taxa. The probes were distributed in the probemixes mostly in an alternating fashion from smallest to largest. Where two consecutive length probes

are included in the same probemix, there is double the minimum length difference between the probes to ensure the sizes can be differentiated without the need for capillary sequencing equipment.

4.3.4 Testing of Probes

A collection of 141 DNA extracts representing 59 Epsilonproteobacterial species and 15 human enteric bacterial pathogens were generated as described in Appendix III. The identity of the DNA extracts from least one strain per species was confirmed using 16S rRNA sequencing as described in Appendix III and DNA extracts were excluded from further analysis if the 16S rRNA-based identification was different to what was expected.

All of the reagents for the MLPA, except the dH_2O, were purchased as SALSA MLPA reagent kits from MRC-Holland (Amsterdam, Netherlands). These kits include positive control DNA and have been optimised for the enzymes and very little can be further optimised. MLPA kits generally recommend a hybridisation time of 16 h but this is impractical for clinical application so a 2 h hybridisation, as employed for an earlier MLPA assay (Cornelius et al. 2014), was used.

MLPA involves four main steps: denaturation of the sample DNA, hybridisation of the LPO, SP (if used) and RPO probes to the sample DNA, ligation of the probes to form single molecules and then amplification of the MLPA products using PCR primers common to all probes. Sample DNA (5 µL) was denatured at 98°C for 5 min and then held at 25°C. Equal volumes (1.5 µL each) of A098 Epsilo A and A099 Epsilo B were added to the denatured sample DNA and the mixture heated at 95°C for 1 min and then held at 60°C for 2 h to facilitate probe hybridisation. The temperature of the tubes was lowered to 54°C and 32 µL Ligase-65 mastermix containing 25 µL dH_2O, 3 µL Ligase Buffer A, 3 µL Ligase Buffer B and 1 µL Ligase-65 enzyme was added to each tube and then the tubes were held at 54°C for 15 min to allow for ligation of the probes followed by heating for 5 min at 98°C to deactivate the enzyme before returning the tubes to 25°C. Polymerase mastermix containing 7.5 µL dH_2O, 2 µL SALSA PCR

primer mix (with 6-carboxyfluorescein [FAM] labelled forward primer) and 0.5 µL SALSA polymerase was added to each tube and then the ligated probes were amplified using 35 cycles of 95°C for 30 sec, 60°C for 30 sec and 72°C for 60 sec followed by 20 min at 72°C before returning the tubes to 25°C. The MLPA products were diluted 1:10 in dH_2O and separated using the ABI genetic analyzer 3130XL (ThermoFisher, Waltham, MA) with POP-7 polymer and the GeneScan 600 LIZ size standard (ThermoFisher). GeneScan results were imported into BioNumerics v7.6.1 and analysed using the amplified fragment length polymorphism (AFLP) module. In order to identify the optimal settings for the band matching module in BioNumerics, the band matching results for a set of 343 Epsilonproteobacteria MLPA reactions were generated using a range of optimization and position tolerance settings. These positive (1) and negative (0) results for each optimization and position tolerance combination were exported as comma-separated values (csv) and saved as a separate worksheet in an Excel (2013; Microsoft, Redwood, WA) workbook. An additional worksheet was generated that had the expected result for each probe in each sample. The worksheets were loaded into R (v 3.2.5, (R Core Team 2016)) using a custom script generated by Dr David Wood and the result for each probe, in each sample and each setting combination was converted to an interpreted result based on the observed and expected results. True positives (TP) had expected and observed band match results of 1, true negatives (TN) had expected and observed band match results of 0, false positives (FP) were 1 for observed and 0 for expected, and false negatives (FN) were 0 for observed and 1 for expected. The number of TP, TN, FP and FN were calculated for each optimization and position tolerance combination. The receiver operating characteristics (ROC) sensitivity and specificity were calculated for each optimization and position tolerance combination. Sensitivity is the TP over the number that should be positive (the sum of TP and FN). Similarity, specificity is the TN over the number that should be negative (the sum of TN and FP). Excel was used to perform two-graph receiver operator characteristic (TG-ROC) analysis which plots the sensitivity and specificity results against each optimization and position tolerance setting.

The specificity of the MLPA probes was evaluated by analysing each of the control DNA extracts individually. DNA was generally used at a concentration of 20 ng/µL but lower concentrations were used if the original extract was lower than this concentration and where the effect of the EDTA was being evaluated. The detection limit of the MLPA assay was evaluated by preparing decimal dilutions of DNA extracts in high EDTA-TE buffer (10 mM Tris [pH 8.0], 1 mM EDTA, in-house) or low EDTA-TE buffer (10 mM Tris [pH 8.0], 0.1 mM EDTA; 12090-015, Invitrogen) and testing these using the Epsilonproteobacteria MLPA assay.

4.4 Results

4.4.1 Pan-genomic Analysis

Thirty-five (35) taxa within Epsilonproteobacteria that are associated with human gastro-intestinal disease or phylogenetically related to *C. jejuni* were identified as possible targets for MLPA development (Table 8). This included three genera, 23 species and nine subspecies, genomospecies or genetically related strains.

Table 8: Taxa Identified as Possible Targets for Epsilonproteobacteria MLPA Assay

Target Taxa	Reason for inclusion	Reference
Arcobacter genus	Genus containing human enteric pathogens	
A. butzleri	Associated with human gastroenteritis and septicaemia	(Collado and Figueras 2011)
A. cryaerophilus	Associated with human gastroenteritis and septicaemia	(Collado and Figueras 2011)
Campylobacter genus	Genus containing human enteric pathogens	
C. avium	Related to *C. jejuni*	(On and Cornelius 2016)
C. canadensis	Related to *C. jejuni*	(On and Cornelius 2016)
C. coli	Associated with human gastroenteritis, related to *C. jejuni*	(On and Cornelius 2016)
C. concisus	Associated with human gastroenteritis and irritable bowel disease	(On and Cornelius 2016)
C. concisus GS1	Genetically distinct group of species associated with human gastroenteritis	(Aabenhus et al. 2005, Ismail et al. 2012, Istivan et al. 2004, Matsheka et al. 2002, Miller et al. 2012, Vandamme et al. 1989)
C. concisus GS2	Genetically distinct group of species associated with human gastroenteritis	(Aabenhus et al. 2005, Ismail et al. 2012, Istivan et al. 2004, Matsheka et al. 2002, Miller et al. 2012, Vandamme et al. 1989)
C. cuniculorum	Related to *C. jejuni*	(On and Cornelius 2016)
C. helveticus	Related to *C. jejuni*	(On and Cornelius 2016)

Target Taxa	Reason for inclusion	Reference
C. hyointestinalis	Associated with human gastroenteritis, related to *C. jejuni*	(On and Cornelius 2016)
C. hyointestinalis subsp. *hyointestinalis*	Subspecies of species associated with human gastroenteritis, related to *C. jejuni*	(On and Cornelius 2016)
C. hyointestinalis subsp. *lawsonii*	Subspecies of species associated with human gastroenteritis, related to *C. jejuni*	(On and Cornelius 2016)
C. insulaenigrae	Associated with human gastroenteritis and septicaemia, related to *C. jejuni*	(On and Cornelius 2016)
C. jejuni	Associated with human gastroenteritis and septicaemia	(On and Cornelius 2016)
C. jejuni subsp. *doylei*	Associated with human gastroenteritis, septicaemia and gastritis	(On and Cornelius 2016)
C. jejuni subsp. *jejuni*	Associated with human gastroenteritis, septicaemia and polyneuropathies	(On and Cornelius 2016)
C. lanienae	Related to *C. jejuni*	(On and Cornelius 2016)
C. lari	Associated with human gastroenteritis, related to *C. jejuni*	(On and Cornelius 2016)
C. lari subsp. *concheus*	Subspecies of species associated with human gastroenteritis	(On and Cornelius 2016)
C. lari subsp. *lari*	Associated with human gastroenteritis, septicaemia and	(On and Cornelius 2016)
C. peloridis	Related to *C. jejuni*	(On and Cornelius 2016)

Target Taxa	Reason for inclusion	Reference
C. subantarcticus	Related to *C. jejuni*	(On and Cornelius 2016)
C. upsaliensis	Related to *C. jejuni*	(On and Cornelius 2016)
Urease-positive thermophilic *Campylobacter* (UPTC)	Related to *C. jejuni* and genetically related to *C. lari*	(Miller et al. 2014)
C. ureolyticus	Associated with human gastroenteritis, wound infections and urethritis	(On and Cornelius 2016)
C. volucris	Related to *C. jejuni*	(On and Cornelius 2016)
Helicobacter genus	Genus containing human enteric pathogens	
H. canis	Associated with human gastroenteritis	(On et al. 2005)
H. cinaedi	Associated with human gastroenteritis	(On et al. 2005)
H. fennelliae	Associated with inflammation of the human rectum and colon	(On et al. 2005)
H. pullorum	Associated with human gastroenteritis	(Burnens et al. 1994, Ceelen et al. 2005, Steinbrueckner et al. 1997)
H. pylori	Human gastric pathogen	(On et al. 2005)

All relevant genomes available on the 24th December 2014 were included in the Epsilonproteobacteria, *Arcobacter, Campylobacter, Helicobacter* and *C. concisus* LS-BSR analyses. It was later recognised that the genome collection included exact duplicates for several *C. lari* genomes, but with different names. In addition, four genomes from UPTC strains were available. Although the taxon is not validly described, they are closely related to *C. lari* (Miller et al. 2014). The six genomes included in the *C. lari* LS-BSR analysis represented four UPTC strains and one strain each of *C. lari* subsp. *concheus* and *C. lari* subsp. *lari*.

During the first stage of genome shortlisting, one genome per strain of all validly described, naturally occurring (so not intentionally modified in the laboratory) strains were generally included. In addition to the exception made for UPTC genomes, the following three exceptions were made. All of the available genomes for genera other than *Arcobacter, Campylobacter* and *Helicobacter* were included. Two draft genomes were shortlisted for the *C. jejuni* type strain because the genome with fewer contigs had higher than expected Prokka-annotated genes. Two genomes for the well characterised *C. jejuni* strain NCTC 11168 were also included in the shortlist, as one was the original complete genome for this species (Parkhill et al. 2000) and the other, also a complete genome, is considered to more closely resemble the original isolate (Revez et al. 2012).

The *Arcobacter* LS-BSR analysis, undertaken on 28 genomes, generated 43,217 predicted coding sequences (CDS, called centroids by LS-BSR). The *Campylobacter* LS-BSR analysis generated 66,235 CDS using 385 genomes and the *Helicobacter* LS-BSR analysis generated 61,991 CDS using 496 genomes. For the smaller LS-BSR analyses, *C. concisus* included 17 genomes and generated 6,208 CDS and *C. lari* included 6 genomes and generated 2,707 CDS.

Genomes were ignored in the initial screening step because they were from species not validly described by the 24th December 2014; alternative genomes of better quality were available for the strain; the strain had been genetically modified; or the strain was subjected to unusual stresses (eg. post-recovery from an animal model). The number of genomes ignored from the *Arcobacter, Campylobacter* and *Helicobacter* LS-BSR results were 7 (25%), 40 (10.4%) and 44 (8.9%), respectively. All of the genomes included in the *C. concisus* and *C. lari* LS-BSR analyses were also used in the initial screening step.

The 17 *C. concisus* genomes were separated into genomospecies 1 (GS1) and genomospecies 2 (GS2) using an alignment of the CON1 and CON2 primers described in Bastyns et al. (1995) with the 23S rRNA genes extracted from the genomes as illustrated in Figure 20.

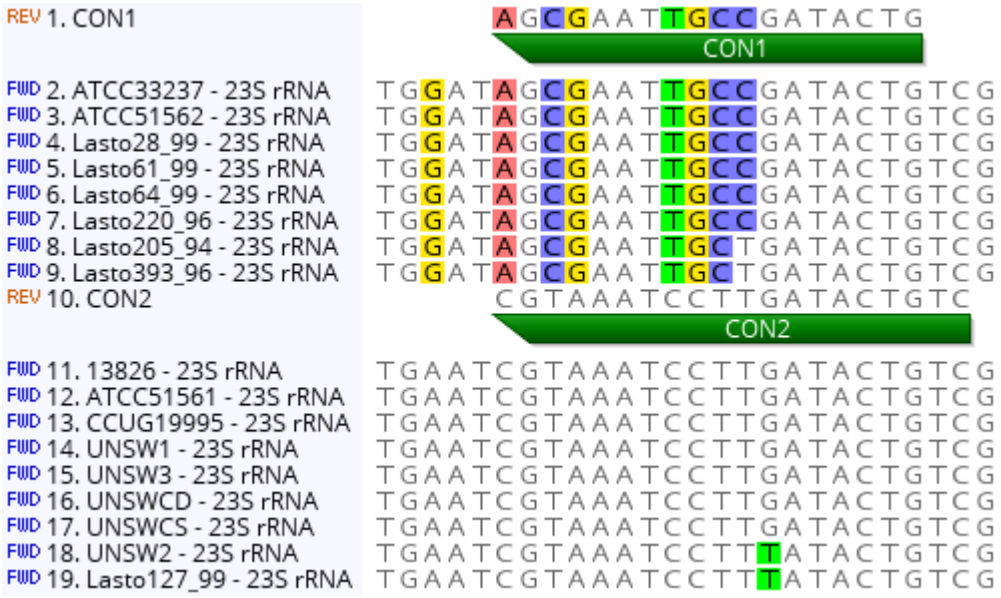


Figure 20: Geneious Alignment of the CON1 and CON2 Primers and 23S rRNA Gene Sequences Extracted from 17 *C. concisus* Genomes

Target (listA) and non-target (listB) genomes were collated from these shortlisted genomes for each possible MLPA target (see Supplementary File 2). The lengths of the CDS were evaluated and a subset of possible taxon-specific CDS was used as queries for searching the local EpsiloFsa BLAST database and the NCBI non-redundant (nr) database using Geneious (R8.1.6 or R9.1.7). Word sizes of 7 or 11 were used and the maximum number of hits was set to 1,000 and 10,000 for EpsiloFsa and nr, respectively.

The `compare_bsr.py` script associated with LS-BSR was successful at identifying CDS suitable for MLPA design for three (*C. helveticus, H. cinaedi* and *H. pullorum*) of the 35 target taxa listed in Table 8. Suitable CDS were identified using the `listAandB.pl` and genus-level LS-BSR analysis for 22 taxa (*Arcobacter* genus, *A. butzleri, A. cryaerophilus, C. avium, C. canadensis, C. concisus, C. cuniculorum, C. hyointestinalis, C. hyointestinalis* subsp. *hyointestinalis, C. hyointestinalis* subsp. *lawsonii, C. insulaenigrae, C. jejuni* subsp. *doylei, C. lanienae, C. lari* subsp. *concheus, C. lari* subsp. *lari, C. peloridis, C. subantarcticus, C. upsaliensis, C. ureolyticus, C. volucris, H. canis* and *H. fennelliae*) and suitable CDS for two additional taxa (*C. concisus* GS2 and UPTC) were identified using `listAandB.pl` and the *C. concisus* and *C. lari* LS-BSR analyses. A small number of *C. concisus* GS1-specific CDS were

identified by `listAandB.pl` but a suitable site for MLPA probe design was not identified because the intra-genomospecies sequence variation was high. The profile-based homology scoring method delta-bitscore (DBS (Wheeler et al. 2016)) identified suitable genes for *C. concisus* GS1. Lasto61.99 and UNSW2 were used as representatives of GS1 and GS2, respectively, and orthologous genes shared by these genomes were identified using best reciprocal phmmer (Finn, Clements, and Eddy 2011) matches. DBSs were generated for each of these orthologous genes in each of the 17 *C. concisus* genomes. The significance of the differences between GS1 and GS2 results were evaluated using a false detection rate (FDR) corrected z-test on data trimmed to remove DBS values >4 standard deviations from the mean, the Kolmogrov-Smirnoff and Mann-Whitney U tests (using FDR corrected scores). Seven genes were present in all 8 *C. concisus* GS1 genomes, had DBS that favoured the *C. concisus* GS1 genomes and had significant scores for each of the tests. Lasto61_99_00770, a putative DNA repair protein, had the highest DBS and was chosen for MLPA development. The bitscores for this gene in the GS1 and GS2 genomes are illustrated in Figure 21. Bitscores are statistical indicators of sequence similarity that are independent of query sequence length and database size and are normalised based on the raw pairwise alignment score[54]. DBS was not required for *C. concisus* GS2 since MPLA probes were able to be designed using CDS identified using listAandB.pl and the *C. concisus* LS-BSR matrix.

[54] https://www.biostars.org/p/187230/

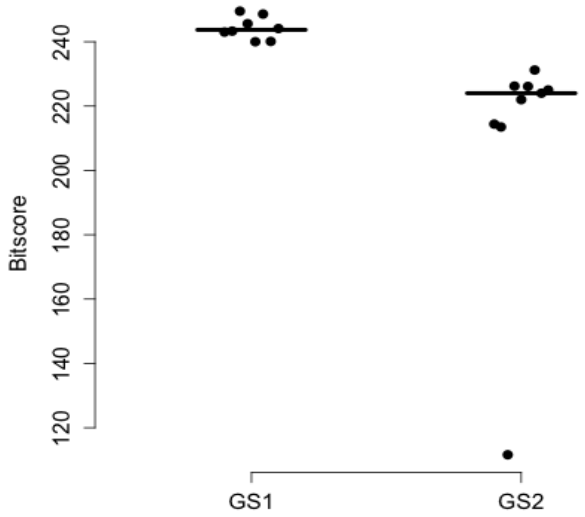


Figure 21: Bitscores for the Lasto61_99_0770 Gene and Orthologous Genes from *C. concisus* Genomes

A genome (10_1_50) recorded as *Campylobacter* sp. in NCBI recorded grades in excess of 95% for all of the *C. concisus* GS2-specific CDS longer than 500 nt that had putative functions annotated by Prokka. Grade is a percentage calculated by Geneious by combining the query coverage, e-value and identity with weights of 0.5, 0.25 and 0.25, respectively (Biomatters Ltd 2015, 194) and aides in the identification of the longest, highest identity hits. To investigate 10_1_50 further, a Geneious (R6.1.7) alignment of the CON1 and CON2 primers, reported by Bastyns et al. (1995), and the 23S rRNA genes extracted from 10_1_50 and the 17 *C. concisus* genomes was generated. The 10_1_50 23S rRNA sequence showed 100% identity with the CON2 primer, as illustrated in Figure 22, and was therefore considered a *C. concisus* GS2 for MLPA probe design.

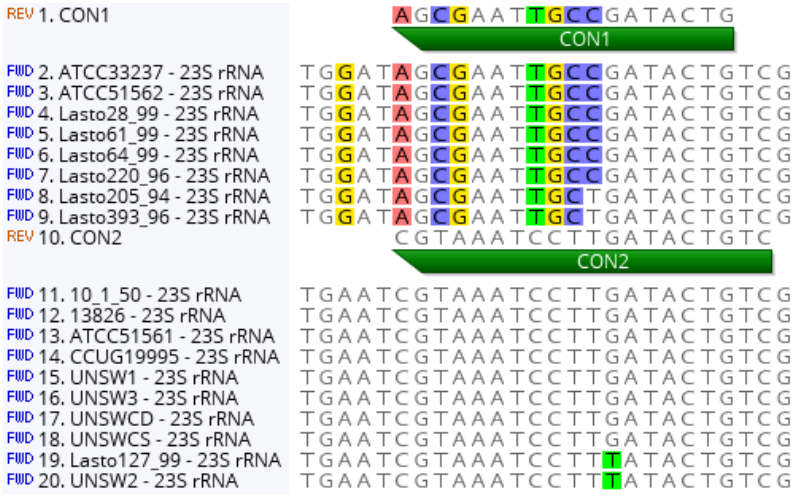


Figure 22: Geneious Alignment of the CON1 and CON2 Primers and 23S rRNA Gene Sequences Extracted from 17 *C. concisus* Genomes and the *Campylobacter* sp. Genome 10_1_50

`Compare_bsr.py` and `listAandB.pl` failed to identify suitable CDS for 7 target taxa (*Campylobacter* genus, *C. coli, C. jejuni, C. jejuni* subsp. *jejuni, C. lari, Helicobacter* genus and *H. pylori*). *C. lari* was abandoned as a target taxon as suitable genes had been identified for the two subspecies and closely related UPTC. *H. pylori* was abandoned as a target taxon as it was a gastric rather than enteric pathogen and so was less relevant for a human gastroenteritis screening panel. Three PCR assays targeting the hippuricase (*hip*O) gene of *C. jejuni* were shown to have 100% specificity and sensitivity when tested with 25 DNA extracts representing 15 *Campylobacter* species including *C. jejuni*, close phylogenetically related species and the type species, *C. fetus* (On, Brandt, et al. 2013). The BLASTn searches using the primers for these PCR demonstrated significant similarity with the gene annotated by Prokka as *yxe*P with the product putative hydrolase YxeP. Five genomes (81-176-DRH212, 81-176-UNCW7, BH-01-0142, BIGS0014 and LMG 23216) did not have this gene annotated and did not show significant similarity with any of the *hip*O PCR primers. The Prokka-annotated gene *yxe*P was the first choice for *C. jejuni* MLPA probe design. Attempts to identify a suitable gene to specifically detect *C. jejuni* subsp. *jejuni* were abandoned once suitable genes for the detection of both *C. jejuni* and *C. jejuni* subsp. *doylei* were identified.

Four *C. coli* PCR targeting different genes demonstrated optimal specificity and sensitivity (On, Brandt, et al. 2013). None of the nine primers and probes showed significant similarity to all 77 *C. coli* genomes. The *C. coli* primers reported by Kawasaki et al. (2008), which target the *gyr*B gene, showed significant similarity to almost all *C. coli* genomes with 76 genomes showing significant similarity with the forward primer and 75 showing significant similarity with the reverse primer. Several other *Campylobacter* species, including some *C. jejuni* strains, had significant similarity to these primers. Similarly, 74 *C. coli* genomes showed significant similarity with both of the *C. coli* primers reported by Wong et al. (2004), which targets the *ceu*E gene, and another two *C. coli* genomes showed significant similarity with the only the reverse primer. Six *C. jejuni* genomes had significant similarity to the reverse primer and one to the forward primer. This *ceu*E gene was the first choice for *C. coli* MLPA design because there were fewer non-*C. coli* genomes with similarity to the primers for this gene.

Both the 16S rRNA and 23S rRNA genes have been reported previously as targets for PCR designed to detect the *Campylobacter* genus (Linton, Owen, and Stanley 1996, Wang et al. 2002) and both showed acceptable specificity and sensitivity values using a limited range of *Campylobacter* species (On, Brandt, et al. 2013). The similarity for the *Helicobacter* genus was too low for the design of MLPA probes using either rRNA gene. As an alternative, the scope of the final target taxon was reduced to those *Helicobacter* species associated with the intestine of humans and animals. This group, which I have called eHelicobacter for the purpose of MLPA design, includes *H. bilis, H. canis, H. cinaedi, H. fennelliae, H. hepaticus, H. pametensis* and *H. pullorum*. `listAandB.pl` identified three CDS that may potentially detect all seven eHelicobacter species but the BLAST results suggested that any probe designed would also detect other *Helicobacter* species and potentially less closely related taxa. Since data on the 23S rRNA had already been generated for the *Campylobacter* and *Helicobacter* genera, this gene was chosen for the design of MLPA probes targeting the *Campylobacter* genus and eHelicobacter.

4.4.2 Design and Manufacture of MLPA Probes

Design of MLPA probes was attempted for each of the 32 shortlisted target taxa as summarised in Table 9 using the criteria provided by MRC-Holland. The MLPA probe names have shown in the font *Eras Light ITS* without periods and spaces between genus initial, species name and subspecies designations e.g. the probe for *C. lari* subsp. *lari* is shown as Clarilari. The specificity of the designed MLPA probes were evaluated by using the probes as queries for BLASTn searches against the local EpsiloFsa database as well as the NCBI nr database using a word size of 7. This probe searches used a smaller word size than was used for the CDS to ensure even small regions of sequence similarity would be detected. MLPA probes were designed for 28 of the target taxa with probes to *C. hyointestinalis, C. hyointestinalis* subsp. *hyointestinalis, C. hyointestinalis* subsp. *lawsonii* and *C. lanienae* not being completed in time for MLPA manufacture within this project.

The *ycl*Q (*ceu*E) gene was extracted from all of the *C. coli* and *C. jejuni* genomes to facilitate the design of a probe specific for the *C. coli* version of this gene. A MLPA probe was designed for *C. coli* version of the *ycl*Q (*ceu*E) gene.

Standard two oligonucleotide (LPO and RPO) MLPA probes, with a single ligation site, were able to be designed for the majority of target taxa. Two probes, Cjejuni and eHelicobacter, were constructed as three oligonucleotides with the middle, spanning probe (SP), introducing a second ligation site.

Table 9: Summary of the Design of Epsilonproteobacteria Multiplex Ligation-dependent Probe Amplification (MLPA) Probes

Taxa	Probe	Selection method[a]	Gene/ centroid	Prokka annotation (similarity to Pfam protein domains)	Reference genome(s)	LHS sequence	SP sequence	RHS sequence
Arcobacter genus	Arcobacter	list Arco	29240	*rps*E, 30S ribosomal protein S5	RM14013 & DSM7299	a CAAGAAGCAATCGTTAAAATCGGAAGAGTAACA		a AAAGTTGTAAAAGGTGGTAGAAGATTCAGATTTACAGCTTTAGTTG
					7h1h	b CAAGAAGCAATCGTTAAGATCGGTAGAGTAACA		b AAAGTTGTAAAAGGTGGAAGAAGATTCAGATTTACAGCTTTAGTTG
A. butzleri	Abutzleri	list Arco	1061	*hcp*, putative beta-lactamace HcpC precursor	7h1h	GGACAAGATGCTTATGATAGACAAGATTATCCTACTGCCATA		AAGTTTTATGAAAAAGCGGCAAGCAAAAACAATGTTGATGCTATTTGGTCTCTTG
A. cryaerophilus	Acryaerophilus	list Arco	53581	hypothetical protein (homologous to Pfam domains PF13384.1 [Homeobox genes – Hox - HTH_23] & PF10668.4 [Phage terminase small subunit])	RM1582	GGAAAAATGATGAAAATAAATCAAAAGCAATAACTATAGTTCAACAAGCAATTA		GAAACGAGTTTGAAGCCCAATATTTTATAGATAATAATCAATTATATACAC
Campylobacter genus	Campylobacter	PCR	23S rRNA	23S rRNA	03-427	GTTCGGTCCCTATCTGCCGTGGGCGTAA		a GAAGATTGAGGAGAGTTGACCCTAGTACGAGAGGACCG
					RM8641			b GAAGATTGAAGAGATTTGACCCTAGTACGAGAGGACCG
C. avium	Cavium	list Camp	626037	type VI secretion protein IcmF	RM8639	CACAGTACTCAGAGGCTAAATGGGCTGAA		ACCATGCTTAGCTCATACCCTAAGACTACTATTGAAATG
C. canadensis	Ccanadensis	list Camp	630135	*saa*, serotype-specific antigen 1 precursor	RM9173	GATGCCTAGTATTGACGCACTTTCAAACAA		AGGTGTAAGATTTTTTAATCAATCCTTTGCTCCAAATG
C. coli	Ccoli	PCR	*ceu*E	*ycl*Q, putative ABC transporter solute-binding protein YclQ precursor	7--1	a GCGTTTGTTACTTCTAAAATCATGGTTTTTAATGATTCTAAGCCATTG		CCACTTGCTAGATACCAGTATTCAGGATCAAGATAAATGATTTT
					BIGS0001	b GCGTTTGTTACTTCTAAAATCATGGTTTTTAGTGATTCTAATCCATTG		
					BIGS0003	c GCGTTTGTTATTTCTAAAATCATGGTTTTTAAAGATTCTAAACCATTG		
					BIGS0006	d GCGTTTGTTATTTCTAAAATCATGGTTTTTAAGGATTCTAAACCATTG		

Taxa	Probe	Selection method[a]	Gene/ centroid	Prokka annotation (similarity to Pfam protein domains)	Reference genome(s)	LHS sequence	SP sequence	RHS sequence
C. concisus	Cconcisus	list Camp	378541	hypothetical protein (no homologous domain in Pfam)	CCUG19995	a CCATCCAAAGGCGTGCCTATAAAAAAGA		a GGGCATTCAAAGATTTGCTACAAATTTCACG
					13826 & RM7084	b CCATCCAAAGGCGTGCTTATAAAAAAGA		b GGGCATTCAAAGATTTGCTGCAAATTTCACG
					Lasto220.96	c CCATCCAAAGGCGCGCTTATAAAAAAGA		
C. concisus GS1	CconcisusGS1	DBS	Lasto61_99_00770	putative DNA repair protein	Lasto61.99	a CCCAAAAAGCCTAAGCATTGCCGAGTTTTATAAAAAGATA		a GTTTATGTAGATGGTAGGTTTGAGATTGATAGCACCTATGC
					ATCC51562 & RM7084	b CCCAAAAAGCCTAAGTATTGCCGAGTTTTATAAAAAGGTA		b GTTTATGTAGATGGTAGATTTGAGATTGATAGCACCTATGC
					Lasto393.96	c CCCAAAAAGCCTAAGTATCGCTGAGTTTTATAAAAAGATA		
					Lasto64.99	d CCCAAAAAGCCTAAGTATCGCCGAGTTTTATAAAAAGATA		
					Lasto28.99	e CCCAAAAATCCTAAGCATTGCCGAGTTTTATAAAAAGATA		
					Lasto220.96	f CCCAAAAAGCCTAAGTATTTCTGAGTTTTATAAAAAGGTA		
C. concisus GS2	CconcisusGS2	list conc	6340	asparate racemace	CCUG19995 & 13826	GCCAGCTTTGACGCCTTTGTAGATGCACTCCATGATGACA		a TCTTGCGTCTCTTTGCTAAAATCCACGCTCTTTAGCCCGCGCTC
					UNSW1			b TCTTGCGTCTCTTTGCTAAAATCGACGCTCTTTAGCCCGCGCTC
C. cuniculorum	Ccuniculorum	list Camp	410795	hypothetical protein (no homologous domain in Pfam)	DSM23162	CACCCCCGACCAAGAAGCTAATAAAGCCGTAA		AAAGTTTTATTGAAAGCGAGATAAAAAGAGCTAAAGAGTATCG
C. helveticus	Chelveticus	comp Camp	583974	hypothetical protein (no homologous domain in Pfam)	RM3228	CCAGAATTGATGCTAGAGTGGGGACTA		TTTTAAACCATTTGTTAAAATCAATTTACCCTATATCTTCCTG
C. hyointestinalis		list Camp	8 possible					
C. hyointestinalis subsp. *hyointestinalis*		list Camp	8 possible					
C. hyointestinalis subsp. *lawsonii*		list Camp	22 possible					

Taxa	Probe	Selection method[a]	Gene/ centroid	Prokka annotation (similarity to Pfam protein domains)	Reference genome(s)	LHS sequence	SP sequence	RHS sequence
C. insulaenigrae	Cinsulaenigrae	list Camp	615634	putative type I restriction enzymeP M protein	NCTC12927	GTGCGTTGCGGTTTTGCTTATCGCTTAA		AAGCCCCTCTGTTTCACTCAAGCCACTTTTGC
C. lanienae		list Camp	599094	putative ATP-binding protein involved in virulence	RM3663			
C. jejuni	Cjejuni	PCR	*hip*O	*yxe*P, putative hydrolase YxeP	NCTC11168 LMG23263	CAAGGCGAATTTGAAAAAATTCGTCATCAA	ATTCATGAAAATCCTGAGCTTGGTTTTGATGAA	a TTATGTACTGCAAAATTAGTGGTGCAAAAATTAAAAGAATTTGb TTATATACTGCAAAATTAGTGGCGCAAAAATTAAAAGAATTTG
C. jejuni subsp. *doylei*	Cjejunidoylei	list Camp	105005	hypothetical protein (similarity to Pfam domain DUF2972, protein of unknown function)	269.97	CATTTAAAACGAAAGGCAATAAACAAGTAATCGGACAATAA		ATAACTTGATTATAGCTATGTTGATTTTCATTTAGGTTAAAGCCGA
C. lari subsp. *concheus*	Clariconcheus	list Camp	579302	N-glycosyltransferase	LMG11760	CCACACTGTGCTTGGACTTTTTTAGCTATGATTTTTGGGAAA		TGTTTTTATTTTAGTGAGCCATTAACTATAGAAAAAAATTTGTGTCCAAG
C. lari subsp. *lari*	Clarilari	list Camp	575534	*via*A, VWA domain protein interacting with AAA ATPase	RM2100	GCAAAATCATAAAAATAAAGGAGCTATAATAATTTGTGTAGATAG		TAGTGGATCCATGGATGGCCAGCCTGAAATCATCGCT
C. peloridis	Cpeloridis	list Camp	483611	exonuclease V alpha subunit	LMG23910	GGATGCTGTTATTGTATTTGAGCAACTTAATGGA		CAATTTTATAGCATAACCTGTAAAATAATAGAAAACATTTTATATTCTG
C. subantarcticus	Csubantarcticus	list Camp	624343	hypothetical protein (no homologous domain in Pfam)	LMG24377	CAATATGGAAATCCAAGTTATGTTATGGGTAACAACA		GTTTATGGGATGGGGAAAATTCGATTGCTAAG
C. upsaliensis	Cupsaliensis	list Camp	411923	*hae*IIM, modification methylase	DSM5365	CACTTTTTGAGGCTATAGGAGACTTGCCAAA		AGATTATGATGAAAATATAAGCAATCACACGGGAAGCACACATAAAA
UPTC	ClariUPTC	list Lari	7	*ure*B, urease subunit beta	CCUG22395	GCTTATGGTTTGTCATCATTTAGATAATAAAATCAAAGAAGATGTTG		AATTTGCAGATAGTAGAATTCGTCCTGAAACTATAGCAGCAG
C. ureolyticus	Cureolyticus	list Camp	610795	multidrug resistance protein	RM4126	GGACAATTTGTAAATCCTTATTTGATTAAATTTATACTTATTTTATTTTCA		TCCAACTCTTCAAATTTAGCATTTTTAATGAGTTTTATTTTCAG
C. volucris	Cvolucris	list Camp	488444	type V secretory pathway, adhesion AidA	LMG24379	GAGCAATCAAAAGAAGCTTTTACCGTAAAAGGTAA		ATTAAACAATGTTCAAGCTTTGTATAATCCTAAATTAGATGTAGTAGGAG

Taxa	Probe	Selection method[a]	Gene/ centroid	Prokka annotation (similarity to Pfam protein domains)	Reference genome(s)	LHS sequence	SP sequence	RHS sequence
Enteric *Helicobacter* species	eHelicobacter	PCR	23S rRNA	23S rRNA	ATCC49314 ATCC BAA-847 NCTC12740 MIT98-5489 MRY12-0050	a CGTCGTGCCAAGAAAAGCC TCTAAGTTTAGCTAATGTT b CGTCGTGCCAAGAAAAGCC TCTAAGTTTAACTAATGTT c CGTCGTGCCAAGAAAAGTTT CTAAGTTTAGCTTATATT d CGTCGTGCCAAGAAAAGCT TCTAAGTTTAGCTAATGTT e CGTCGTGCCAAGAAAAGCC TCTAAGTTTAACAATGGTT	GCCCGTACCG CAAACCGACA CAGGTAGA	TGAGATGAGTATTCTAAGGCG CGTGAAAGAACTC
H. canis	Hcanis	list Heli	455723	*cdt*B, deoxyribonuclease CdtB	NCTC12740	GCCTTATGCGACCACACTTCC TATCCA		AGAATACCAATGGGACCTAGG CAGGG
H. cinaedi	Hcinaedi	comp Heli	87786	hypothetical protein (Pfam search found weak similarity to DUF2273 - small integral membrane protein)	ATCC BAA-847	GGGTTACACTTCCCCACATAT TTATAGTAGTGTGGCTTA		AGTCTAAAAACTTTGCTAATTC ATTTTTGTTTTTAAATCCTGC
H. fennelliae	Hfennelliae	list Heli	442759	uracil-DNA glycosylase, family 4	MRY12-0050	GAGCAAACGCACTTATGCAGA CCATATTTGCAAGAGCAGCTT A		GTGGATTGCGCGCAAAAGTTG TGATCGTGTTTGATGCG
H. pullorum	Hpullorum	comp Heli	431322	hypothetical protein (homologous domains in Pfam to PF13302.1 [Acetyltransferase (Hardo et al.) domain - Acetyltransf_3 and PF13474.1 [SnoaL-like domain, SnoaL_3]	MIT98-5489	GGAAAAATTGAATTTTATACA GAAAGGCTAATTATTAGGCAA TTATGC		GAAAATGATCTTGAAGCATATT TCAAACTTCTCAATAACCCAAA AGC

[a] list: `listAandB.pl`; comp: `compare_bsr.py`; DBS: delta-bitscore; PCR: published PCR; Arco: *Arcobacter* LS-BSR; Camp: *Campylobacter* LS-BSR; conc: *C. concisus* LS-BSR; lari: *C. lari* LS-BSR; Heli: *Helicobacter* LS-BSR

All of the probe sequences were checked by MRC-Holland and then the lengths of some probes were modified by the inclusion of non-annealing DNA in the form of small “stuffers” of up to 4 nucleotides and/or M13 vectors of lengths ranging from 58 to 391 nucleotides. This allowed two probemixes to be generated that, when used separately, provide adequate differentiation of all probes using microfluid electrophoresis systems. The two probemixes could also be used in a single reaction when capillary electrophoresis equipment was used for product detection. Table 10 summarises the constituents of the two probemixes A098 Epsilo A and A099 Epsilo B.

Table 10: Summary of Epsilonproteobacteria Multiplex Ligation-dependent Probe Amplification (MLPA) Probemixes

A098 Epsilo A Probemix

Probe	LPO stuffer (nt)	LHS (nt)	SP (nt)	RHS (nt)	RPO stuffer (nt)	Total (nt)
Cupsaliensis		31		47		120
Cureolyticus		51		44	3	140
Cconcisus		28		31		180
Arcobacter		33		46		200
Hcinaedi		39		43		224
Acryaerophilus		54		51	1	248
Abutzleri		42		55		272
Campylobacter	1	28		38		296
Ccoli	1	48		44		322

A099 Epsilo B Probemix

Probe	LPO stuffer (nt)	LHS (nt)	RHS (nt)	RPO stuffer (nt)	Vector (nt)	Total (nt)
Cjejunidoylei	1	41	46		0	130
Chelveticus		27	43		58	170
Hfennelliae		43	38		67	190
Csubantarcticus	1	37	32		100	212
Clarilari		45	37		112	236
Cpeloridis	2	34	49		133	260
Hcanis	2	27	26		187	284
Ccuniculorum	1	32	43		190	308

A098 Epsilo A Probemix

Probe	LPO stuffer (nt)	LHS (nt)	SP (nt)	RHS (nt)	RPO stuffer (nt)	Total (nt)
Cjejuni		30	33	43		350
eHelicobacter		38	28	34	4	378
CconcisusGS1		40		41		406
Hpullorum		48		47		462
CconsisusGS2	1	40		44		518

A099 Epsilo B Probemix

Probe	LPO stuffer (nt)	LHS (nt)	RHS (nt)	RPO stuffer (nt)	Vector (nt)	Total (nt)
Ccanadensis		30	38		226	336
Cavium		29	39	1	253	364
Cinsulaenigrae	2	28	32		316	420
ClariUPTC		47	42	51	316	448
Cvolucris		35	50		349	476
Claricon cheus		42	60		370	504

4.4.3 Testing of Probes

A total of 127 DNA extracts, representing 80 taxa, were tested using the Epsilonproteobacteria MLPA assay (Appendix IV). Early results demonstrated that small "aberrant" peaks, with lengths similar to the probe 10-15 nt shorter, were sometimes being observed in addition to some expected peaks. For example, the DNA from the *A. butzleri* CCUG 30485^{T} was positive for the Arcobacter and Abutzleri probes but also produced a small peak at a length very close to the expected length of the Hfennelliae product (Figure 23).

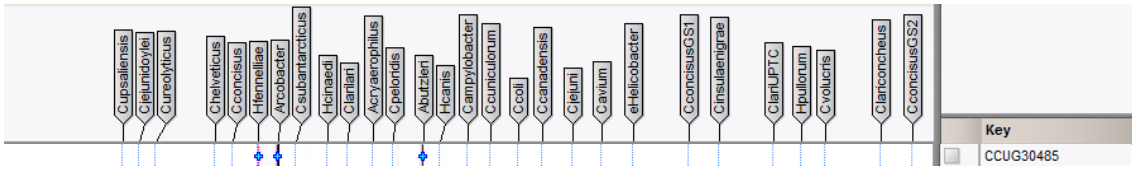


Figure 23: An Example of the "Aberrant" Peaks Observed with the Epsilonproteobacteria Multiplex Ligation-dependent Probe Amplification (MLPA) Assay
The Arcobacter and Abutzleri peaks are expected and the Hfennelliae peak is an "aberrant" of the Arcobacter peak

To minimise the effect of the aberrant peaks, the position tolerance within the band matching module of BioNumerics was varied from 0.03 to 0.20%. The optimization was varied from 0 to 2.0% for a subset of position tolerance settings. The resulting band matches were exported as comma-separated values (csv) and the receiver operator characteristics (ROC) specificity and sensitivity were generated using R. Optimization had no effect on band matching (data not shown). Two-graph receiver operator characteristic (TG-ROC) analysis (Figure 24) demonstrated that the optimal position tolerance setting was 0.13%. This setting was used for all of the 308 MLPA reactions used to test the probes. This includes results for the MLPA positive control, which contains target sequences for all 28 probes, the no template control (low EDTA-TE buffer) and multiple analyses for some DNA extracts including different DNA concentrations. Even with this setting, unexpected peaks were assigned by BioNumerics in 17 MLPA reactions, 8 of which were "aberrant" peaks. Conversely, with the position tolerance set to 0.13%, 131 MLPA reactions had at least one expected peak not assigned. The expected but

unassigned peaks were visible in 61 reactions. These peaks were manually assigned and have been italicised in Appendix IV. Table 11 summarises the results for each of the probes for the 308 MLPA reactions.

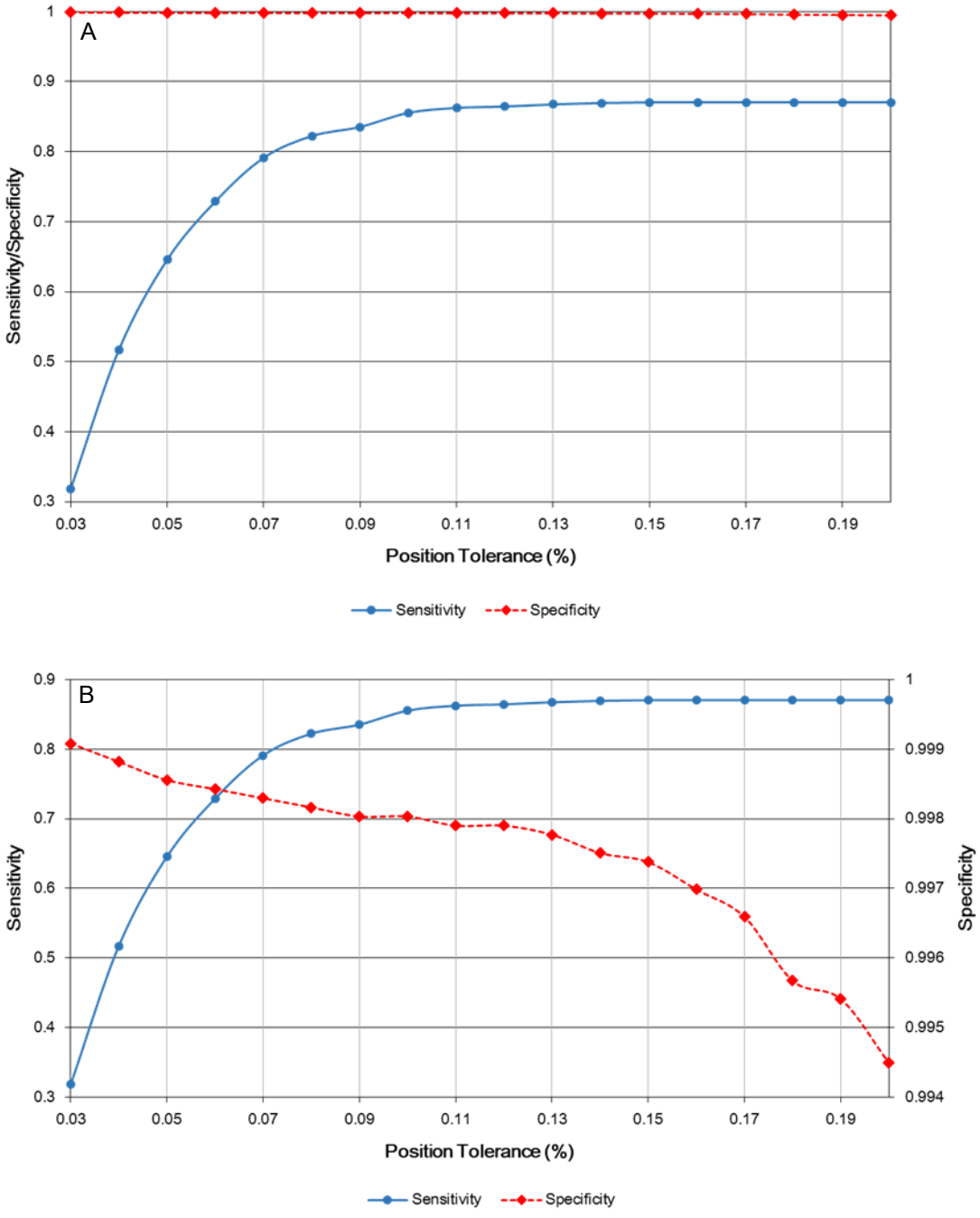


Figure 24: Two Graph Receiver Operator Characteristics (TG-ROC) Analysis for the Epsilonproteobacteria Multiplex Ligation-dependent Probe Amplification Assay
Panel A is the standard format for TG-ROC and Panel B has separate scales so the variation in specificity can be observed.

Table 11: Receiver Operator Characteristics for the 28 Epsilonproteobacteria Multiplex Ligation-dependent Probe Amplification (MLPA) Probes for 308 Reactions

Probes	True Negative	True Positive	False Negative	False Positive
Cupsaliensis	282	22	4	0
Cjejunidoylei	285	22	1	0
Cureolyticus	285	22	1	0
Chelveticus	283	23	2	0
Cconcisus	236	69	3	0
Hfennelliae	283	22	2	1
Arcobacter	216	83	1	8
Csubantarcticus	278	25	1	4
Hcinaedi	285	21	2	0
Clarilari	286	20	2	0
Acryaerophilus	285	21	2	0
Cpeloridis	285	21	2	0
Abutzleri	281	25	2	0
Hcanis	285	21	2	0
Campylobacter	147	118	43	0
Ccuniculorum	282	22	2	2
Ccoli	282	24	2	0
Ccanadensis	284	22	2	0
Cjejuni	278	26	4	0
Cavium	286	20	2	0
eHelicobacter	252	45	11	0
CconcisusGS1	263	36	8	1
Cinsulaenigrae	285	19	4	0
ClariUPTC	284	21	3	0
Hpullorum	284	17	7	0
Cvolucris	285	21	2	0
Clariconcheus	285	18	5	0
CconcisusGS2	259	38	10	1

The expected MLPA probes were detected for the majority (187, 97.9%) of the 260 MLPA reactions. All of the 33 DNA extracts representing 17 taxa in the *Arcobacter* genus produced the expected results with all being positive for the Arcobacter probe and the extracts from *A. butzleri* and *A. cryaerophilus* strains producing peaks corresponding to the Abutzleri and Acryaerophilus probes, respectively.

A total of 62 DNA extracts from 29 *Campylobacter* taxa were tested. Of the 141 MLPA reactions, 86 (61.0%) produced the expected results. DNA extracts from three *Campylobacter* taxa consistently produced unexpected results. All three MLPA reactions for the DNA extract from *C. lari* subsp*. concheus* RM 14091^T were negative for the Clariconcheus probe. All four MLPA reactions for the DNA extract from *C. subantarcticus* strain CCUG 38513^T were positive for the Arcobacter probe. This is likely an "aberrant" peak but the length is indistinguishable from that of the true Arcobacter peak and could not be excluded using the tolerance settings. This was evaluated further in 4.4.4 Additional Investigations below. The four MLPA reactions for the DNA extract from UPTC were all positive for ClariUPTC, as expected, but they were also positive for Csubantarcticus and one of the MLPA reactions was also positive for Arcobacter. The Arcobacter peak may be an "aberrant" from the Csubantarcticus peak. The unexpected Csubantarcticus peak observed with UPTC DNA was evaluated in more depth in 4.4.4 Additional Investigations below.

Half (n = 22) of the 44 reactions with unexpected results for the *Campylobacter* genus not involving *C. lari* subsp*. concheus, C. subantarcticus* or UPTC DNA were negative for the Campylobacter probe including three of the four DNA extracts prepared for each of *C. coli* and *C. jejuni* in low-EDTA TE buffer. In contrast, this probe was detected in all 42 MLPA reactions where the DNA concentration was <20 ng/µL, all of which were prepared in buffers containing <1 mM EDTA.

ACP170b and L395, two of the four DNA extracts from *C. upsaliensis*, were unexpectedly negative for the Cupsaliensis probe. Possible reasons for this result were evaluated in 4.4.4 Additional Investigations below.

The two genomospecies of *C. concisus* failed to consistently produce expected results even when the results for the Campylobacter probe were ignored. Two MLPA reactions for the DNA extracts from the *C. concisus* GS2 strains Lasto104.93 and Lasto131.99 had no target probes detected and no Q-fragments visible. Of the 16 MLPA reactions for GS1 DNA extracts at 20 ng/µL (in high-EDTA TE buffer), five were unexpectedly negative for the CconcicusGS1 probe

and four were also negative for the Campylobacter probe. Likewise, of the 20 MLPA reactions for GS2 DNA extracts at 20 ng/μL (in high-EDTA TE buffer), four were unexpectedly negative for CconcicusGS2, all of which were also negative for the Campylobacter probe. The expected results were observed for 14 of 16 of the MLPA reactions for *C. concisus* DNA extracts diluted in low-EDTA TE buffer. The two reactions that still had unexpected results were for the *C. concisus* GS1 strain Lasto24.99 and the *C. concisus* GS2 strain Lasto104.93 which were positive for the probe specific for the other genomospecies in addition to the three expected probes.

One MLPA reaction for the *C. coli* CCUG 11283^{T} was positive for the Ccuniculorum probe in addition to the expected Campylobacter and Ccoli probes. This is also likely to have been an "aberrant" peak from Ccoli. The MLPA results for the remaining 13 *Campylobacter* taxa were concordant with the expected results, if the Campylobacter probe was ignored.

A total of 18 DNA extracts, representing 17 *Helicobacter* species were tested using the Epsilonproteobacteria MLPA. Both of the MLPA reactions for *H. pullorum* CCUG 33837^{T} at 20 ng/μL were negative for all target probes and no Q-fragments were visible suggesting these reactions were inhibited. When this DNA was diluted to 2 ng/μL using low-EDTA TE buffer, the eHelicobacter probe was detected but the Hpullorum probe was negative. Further investigations into the Hpullorum probe and *H. pullorum* DNA was undertaken in 4.4.4 Additional Investigations below.

The three other *Helicobacter* species for which species-specific MLPA probes were included were all positive for the appropriate species-specific probes and negative for eHelicobacter when tested at 20 ng/μL in high-EDTA TE buffer. All three were positive for eHelicobacter, in addition to the appropriate species-specific probe, when the DNA was diluted to 2 ng/μL in low-EDTA TE buffer before testing. One of the two MLPA reactions for the *H. fennelliae* CCUG18820^{T} at 2 ng/μL was also positive for the Arcobacter probe. This is not likely to be an "aberrant" band since the true positive band is smaller rather than larger than the false positive band. The eHelicobacter probes were designed to detect the enteric *Helicobacter* species *H. bilis, H. canis, H. cinaedi, H. fennelliae, H. hepaticus, H. pametensis* and *H.*

pullorum, but it was expected that additional *Helicobacter* species would also be detected using these probes. The eHelicobacter probes detected eight other *Helicobacter* species (*H. canadensis, H. cholecystus, H. ganmani, H. mesocricetorum, H. muridarum, H. mustelae, H. typhonius* and *'H. winghamensis'*). Two *Helicobacter* species, *H. pametensis* and *H. pylori,* were consistently negative for the eHelicobacter MLPA probe. The DNA from the two *H. pylori* strains was also negative for eHelicobacter when diluted to 2 ng/μL in low-EDTA TE buffer but there was insufficient MLPA probemix available to test diluted DNA from the other two species so it is possible that DNA from *H. pametensis* may test positive for the eHelicobacter probe.

No false negative results were observed for any of the 100 MLPA results for Epsilonproteobacterial DNA (excluding *C. lari* subsp*. concheus, C. upsaliensis* and *H. pullorum*) with a concentration of less than 20 ng/μL. Almost all (n = 98) of these MLPA reactions involved DNA extracts and dilutions that were prepared using dH_2O or TE buffer containing <1 mM EDTA. The resuspension buffer was unknown for remaining two DNA extracts.

Fourteen of the 15 DNA extracts from human enteric bacterial pathogens produced MLPA results with only Q-fragments visible. The DNA extract from the *Clostridium difficile* strain NZRM 2390^T was unexpectedly positive for the Ccuniculorum probe. A review of the BLAST search of the NCBI nr database confirmed that the only hit was to a *C. cuniculorum* genome and there was no hits to *Clostridium difficile* genomes. It is therefore most likely that the *Clostridium difficile* DNA extract had become contaminated with *C. cuniculorum* DNA at some stage in the testing process.

DNA solutions prepared in high EDTA-TE buffer, with the formulation of 10 mM Tris pH 8.0 and 1 mM EDTA, failed to generate any peaks at concentrations below 20 ng/μL (data not shown). New decimal dilution series, from 20 ng/μL to 20 pg/μL, were prepared for DNA extracts from two *C. concisus* strains (CCUG 19995 and Lasto275.95), *C. jejuni* subsp. *jejuni* CCUG 11284^T and *C. avium* RM 8639^T using the low EDTA-TE buffer (12090-015, Invitrogen), which had the formulation of 10 mM Tris pH 8.0 and 0.1 mM EDTA. These dilutions were tested in duplicate (where possible) and on two occasions. Given that replicates c and d were tested

three days after replicates a and b, they could be considered biological replicates. There was an insufficient quantity of *C. avium* RM8639^T to perform duplicates of the highest concentrations. The detection limit of the MLPA varies for different DNA extracts as shown in Table 12.

Table 12: Evaluation of the Detection Limit of the Epsilonproteobacteria Multiplex Ligation-dependent Probe Amplification (MLPA) Assay
Replicates c and d were tested 3 days after replicates a and b

DNA Extract	20 ng/µL				2 ng/µL				200 pg/µL				20 pg/µL			
Probe	a	b	c	d	a	b	c	d	a	b	c	d	a	b	c	d
CCUG11284^T																
Campylobacter	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Cjejuni	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
CCUG19995																
Campylobacter	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-
Cconcisus	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
CconcisusGS2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+
Lasto275.95																
Campylobacter	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-
Cconcisus	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-
CconcisusGS2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-
RM8639^T																
Campylobacter	+				+	+	+	+	+	+	+	+	+	+	+	+
Cavium	+				+	+	+	+	+	+	+	+	+	+	+	+

4.4.4 Additional Investigations

In order to further evaluate whether the Arcobacter peak observed with *C. subantarcticus* DNA was an “aberrant” peak associated with the Csubantarcticus peak, 20 ng/µL and 2 ng/µL concentrations of DNA extracted from *C. subantarcticus* CCUG 38513^T were tested with A098 EpsiloA and A099 EpsiloB probemixes in separate MLPA reactions and the results are shown in Figure 25. Both concentrations were, as expected, positive for Campylobacter using EpsiloA. Both concentrations also produced the expected peak for Csubantarcticus using EpsiloB and were also positive for Arcobacter even though the probes for this target are in EpsiloA and not EpsiloB.

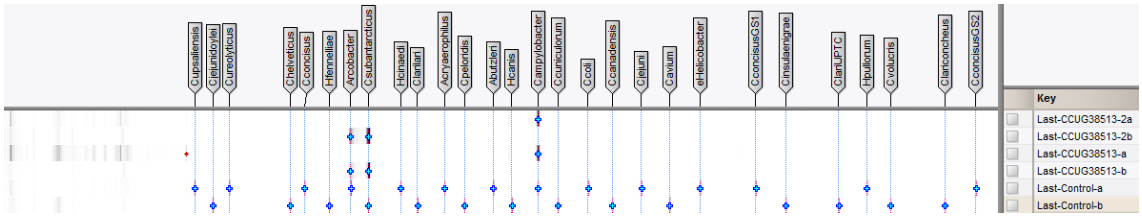


Figure 25: Epsilonproteobacteria Multiplex Ligation-dependent Probe Amplification Results for *C. subantarcticus* Strain CCUG 38513^T using the A098 EpsiloA and A099 EpsiloB Probemixes Separately

The suffix -2 indicated the DNA was at a concentration of 2 ng/µL. The suffix a indicates the EpsiloA probemix and the suffix b indicated the EpsiloB probemix.

A peak consistent with the Csubantarcticus probe was observed with UPTC DNA. *C. subantarcticus* and UPTC both belong to the *C. lari* group (Miller et al. 2014) so some similarity would be understandable. The results of the local EpsiloFsa and NCBI nr BLAST searches (word size 7) for the Csubantarcticus probe were reviewed but none of the four UPTC genomes showed any significant similarity to the Csubantarcticus probe (joined LHS+RHS) and the only hits were to the target genomes. A local BLAST database was generated of all the Epsilonproteobacteria MLPA probes (joined LHS+RHS) in Geneious R9.1.7 and a BLASTn search with a word size of 7 was conducted with the complete genome of the Enterobacteria phage M13 (NC_003287.2) as the query. No hits were returned demonstrating that there was no significant similarity between the Csubantarcticus probe and the phage used to generate the extended RPO.

Half of the *C. upsaliensis* DNA extracts were unexpectedly negative for the Cupsaliensis probe. In order to evaluate the reason for this unexpected result, both the Cupsaliensis probe sequence and the *hae*IIM (modification methylase) gene were used as queries against the local KrunoFsa database that was generated using the genomes from three *C. helveticus* and three *C. upsaliensis* strains, including the genome for the strain ACP170b. Both the probe and gene had significant similarity to two *C. upsaliensis* strains but no significant similarity was found for either the probe or gene for the ACP170b genome.

To evaluate whether the negative results observed for *H. pullorum* DNA was due to high-EDTA in the TE buffer, a new DNA extract for this strain was produced and the DNA eluted and diluted to 20 ng/µL and 2 ng/µL in low-EDTA TE-buffer. Both concentrations of DNA from the new extract were positive for eHelicobacter and negative for Hpullorum.

4.5 Discussion

4.5.1 Epsilonproteobacterial Genomes

Next generation sequencing using platforms such as the Illumina MiSeq has resulted in a dramatic increase in the number of publicly available bacterial whole genomes (Land et al. 2015). This has facilitated the comparisons of bacteria at a genome level for differences that can explain variations in attributes such as pathogenicity (Lukjancenko, Ussery, and Wassenaar 2012, Sahl et al. 2013), host range (Ben Zakour et al. 2012) and to better understand bacterial evolution (Skarp-de Haan et al. 2014). These types of analyses, however, are only as good as the genomes included in the studies. Complete genomes offer the greatest assurance of quality but require more extensive sequencing, often using a number of different sequencing platforms, so are generally reserved for the first genome of a species, often the type strain, and possibly a few key alternative strains of particular interest. Draft genomes provide a cost effective option for generating additional information about a bacterial species and the number of contigs can provide an initial indication of draft genome quality, especially if comparing genomes from the same strain.

A total of 91 genomes were ignored in the initial screening for taxon-specific genes because they were either duplicate genomes, were from intentionally modified strains, were strains identified only to the genus level, or were from taxa that had not been validly described by 24th December 2014.

4.5.2 Pan-genomic Analysis and Target Gene Selection

This study used LS-BSR for pan-genomic analysis of more diverse genomes than previously reported (Baig et al. 2015, Sahl et al. 2015, Sahl et al. 2014, Sahl et al. 2017, Wallace et al. 2016). Each of the genus- and class-level LS-BSR analyses in this study took considerably longer than the original paper which may be due, in part, to the fewer processors being used but for the majority of the time only a single processor was being used. This was especially the case for the full Epsilonproteobacteria LS-BSR analysis which involved 939 genomes and generated over 2 million CDS.

The LS-BSR package included the `compare_bsr.py` script for identifying CDS present in a subset of genomes. This script failed to identify any CDS for most of the genome groups attempted. Given that `listAandB.pl` used the same data as `compare_bsr.py` and was able to identify taxon-specific CDS in most cases, the criteria for identifying group-specific CDS in `compare_bsr.py` must be quite stringent. The `listAandB.pl` developed within the current study was flexible and easy to use. Both `compare_bsr.py` and `listAandB.pl` required lists of target and non-target genomes.

The majority of *C. concisus* strains for which genomes were publicly available did not already have genomospecies assigned so it was necessary to assign temporary genomospecies so that the genomes could be placed into two lists. This was achieved by aligning the reverse primers described by Bastyns et al. (1995) and the 23S rRNA gene sequences extracted from the genomes. Later BLAST searches of possible taxon-specific CDS for *C. concisus* GS2 revealed that *Campylobacter* sp. genome 10_1_50 had significant similarity to all of the candidate CDS for this taxon. A second alignment confirmed that the 23S rRNA gene sequence extracted from this genome had 100% identity with the CON2 primer placing it within the *C. concisus* GS2. The assignment of genomospecies to the 18 genomes, including 10_1_50, was later shown to match the results of additional genomic comparisons reported in Chapter 3: Comparative Analysis of *Campylobacter concisus*.

Delta-bitscore (DBS), a method for identifying divergent orthologous genes, was applied to the *C. concisus* genomes and identified seven genes that might prove useful for identifying *C. concisus* GS1. The gene with the largest DBS was the first choice for MLPA design for this taxon. This method currently compares pairs of genomes and involves the generation of custom hidden Markov models (Wheeler et al. 2016) and is not yet suitable for evaluating large sets of genomes.

Published PCR primers were a useful resource for identifying possible taxon-specific genes for the four taxa where suitable genes were not identified using `compare_bsr.py` or `listAandB.pl`. The *hip*O (*yxe*P) gene was chosen for *C. jejuni* and the 23S rRNA gene was chosen for both *Campylobacter* and eHelicobacter, a narrowed taxon reflecting the difficulty of designing a probe generic enough to detect the entire *Helicobacter* genus.

Attempts to identify a *C. coli* –specific gene within the *Campylobacter* LS-BSR data were unsuccessful and BLAST searches of a selection of published PCR primers for this species failed to find a gene target with high similarity to all 77 *C. coli* genomes and no non-*C. coli* genomes. The failure to identify regions homologous to the PCR primers in some genomes may reflect the inclusion of draft genomes since it has been reported that high-quality draft genomes (length of ≤ 2 Mb and < 150 contigs) from *C. jejuni* and *C. coli* could be missing up to 5% of the 1343 core genes incorporated into v1.0 of the core genome multi-locus sequence typing (cgMLST) scheme for these taxa (Cody et al. 2017). It is also possible that some of the genomes have been misidentified (Bull et al. 2012). Primers to the *gyr*B and *ceu*E genes showed significant similarity for the highest number of *C. coli* genomes and the *ceu*E was the first choice for MLPA design because fewer non-*C. coli* genomes showed significant similarity to these primers.

4.5.3 Design and Manufacture of MLPA Probes

Probes were designed for 28 of the 35 taxa initially identified for MLPA development. Unfortunately, time ran out for the design of probes for four taxa (*C. hyointestinalis*, the two *C.*

hyointestinalis subspecies and *C. lanienae*), *H. pylori* was excluded as it is a gastric rather than enteric pathogen and the two taxa *C. jejuni* subsp. *jejuni* and *C. lari* were abandoned as probes had been designed for sufficiently related taxa to provide adequate identification without these additional probes.

4.5.4 Testing of Probes

The optimization setting within the BioNumerics band matching algorithm did not affect the specificity or sensitivity of the Epsilonproteobacteria MLPA assay. The TG-ROC analysis demonstrated that the specificity was consistently very high (0.994-0.999) and only moderately dependent on the position tolerance setting. Conversely, the sensitivity varied from 0.319 to 0.872 and was very highly dependent on the position tolerance setting. Ordinarily, the optimal setting is the intersection of the two graphs, where specificity and sensitivity are equal (Greiner 1995). The two graphs within the TG-ROC for the Epsilonproteobacteria MLPA assay did not cross so the position tolerance (0.13) that maximised both the sensitivity and specificity was chosen for subsequent analyses.

Over half (15/28, 53.6%) of the Epsilonproteobacteria MLPA probes consistently produced the expected results. The probe for Ccoli should work as well as any published PCR for differentiating *C. coli* from *C. jejuni*. The remaining 14 probes that produced results concordant with expectations were Abutzleri, Acryaerophilus, Cavium, Ccanadensis, Chelveticus, Cinsulaenigrae, Cjejunidoylei, Cjejuni, Clarilari, Cpeloridis, Cureolyticus, Cvolucris, Hcanis and Hcinaedi.

Two probes, Clariconcheus and Hpullorum, failed to produce the expected peaks for DNA extracted from strains of their target taxa. Genomes were available for only one strain for each of these taxa and the DNA extracts were only available from single strains that differed from those for which genomes were available. It is therefore likely that the genes identified by LS-BSR were part of the accessory genomes of these taxa, rather than the core. Several other possible taxon-specific genes were identified for both of these taxa. Alternative probes could be

designed to replace the current ones but additional genomes and strains are required to adequately identify appropriate taxon-specific genes and test the specificity of the resulting probes.

Both replicates of the *H. pullorum* DNA extract at 20 ng/μL and one replicate each of two *C. concisus* DNA extracts at the same concentration produced no peaks with the Epsilonproteobacteria MLPA including no Q-fragments. There was no identifiable differences in the methods used to prepare these extracts, relative to remaining extracts. Q-fragments are included at a low concentration as part of the probemixes and are designed to be detected only when target DNA quantity is below 100 ng (MRC-Holland 2015). Failure to detect these Q-fragments when no other probes are detected suggests that these reactions were inhibited. False-negative results caused by inhibition could result in the incorrect assumption that a sample is free of a pathogen (Cone, Hobson, and Huang 1992). Additional analyses, for example re-extraction or dilution, can be applied to samples exhibiting signs of inhibition in order to improve confidence in the result.

Two of the four *C. upsaliensis* control DNA extracts tested were positive for the Cupsaliensis probe but ACP170b and L395 were unexpectedly negative for this probe. ACP170b was recovered from a cat in New Zealand (Bojanic 2017, Bojanic et al. 2017), and L395, was isolated from a human stool sample in Belgium. A BLAST search for both the probe and the *hae*IIM gene failed to find any similarity with the genome generated from ACP170b. DNA extracts from additional *C. upsaliensis* strains are required to establish the prevalence of this gene in this species. In order to detect all *C. upsaliensis* strains an additional probe will need to be designed and included in the probemix.

Of the 180 MLPA reactions conducted on Epsilonproteobacterial DNA (excluding *C. lari* subsp*. concheus, H. pullorum* and the two *C. upsaliensis* DNA extracts ACP170b and L395) diluted in high-EDTA TE buffer, almost a quarter (44, 24.4%) were negative for at least one expected peak. Three quarters of the DNA extracts prepared for *C. coli* and *C. jejuni* in low-EDTA TE buffer were negative for the Campylobacter probe. Conversely, no false negative

results were observed for any of the Epsilonproteobacterial DNA (excluding *C. lari* subsp*. concheus, C. upsaliensis* and *H. pullorum*) with concentrations less than 20 ng/μL. Most, if not all, of these lower concentration DNA extracts and dilutions were prepared using dH_2O or low EDTA-TE buffer. The initial set of dilutions prepared to test the detection limit of the Epsilonproteobacteria MLPA was prepared in high EDTA TE buffer and only the 20 ng/μL dilutions produced the expected peaks. Conversely, the second set of dilutions was prepared in low EDTA-TE buffer and several dilutions, for each of the four DNA extracts, produced the expected peaks. These results, taken together, suggest EDTA rather than DNA concentrations at or above 20 ng/μL adversely affects the generation of some MLPA products. EDTA has been shown to inhibit PCR at concentrations above 0.1 mM (Al-Soud and Radstrom 2001, Kreader 1996, Rossen et al. 1992) by chelating the magnesium ions (Rossen et al. 1992) which are an essential cofactor of the polymerase enzyme. The EDTA contributed by DNA prepared in high EDTA-TE buffer (1 mM) results in a concentration of 0.1 mM in the MLPA PCR reaction, a concentration not previously shown to be inhibitory. It is possible that the inhibited step in the MLPA procedure is the ligation step. The human DNA Ligase I enzyme has been shown to be magnesium-dependent (Taylor et al. 2011) and the T4 DNA ligase is also sensitive to EDTA[55]. The source of the Ligase-65 enzyme used in the MLPA assay is proprietary but it is possible that it, too, is dependent on magnesium ions. Given that the reaction volume at the ligation step is 40 μL, rather than the 50 μL for the MLPA PCR step, it is possible that this slightly higher EDTA concentration could affect the performance of the ligase resulting in false negative results.

The eHelicobacter probe was designed to detect the seven enteric *Helicobacter* species *H. bilis, H. canis, H. cinaedi, H. fennelliae, H. hepaticus, H. pametensis* and *H. pullorum* although it was recognised that other *Helicobacter* species, and potentially other non-

[55] https://www.sigmaaldrich.com/technical-documents/protocols/biology/roche/rapid-dna-ligation-kit.html accessed 25th August 2018

Helicobacter Epsilonproteobacteria taxa, may have enough similarity to be detected with these probes. Both MLPA reactions containing DNA from the *H. pametensis* type strain were negative for the eHelicobacter probe as were all eight MLPA reactions containing DNA from the two *H. pylori* strains and one of the two MLPA reactions containing DNA from the *H. muridarum* type strain. Based on the similarity of 23S rRNA sequences extracted from the genomes for these species, *H. muridarum* may be detected by the eHelicobacter probe with most *H. pylori* strains less likely to be detected. No non-*Helicobacter* species were positive for eHelicobacter which suggests the second ligation site introduced by the inclusion of a spanning probe (SP), has successfully prevented non-target genera from being detected. This requirement for simultaneous annealing of three parts to facilitate two ligation events may result in this probe being more strongly affected by the presence of EDTA.

Of the 253 MLPA reactions for control DNA (*C. lari* subsp*. concheus* and *H. pullorum* excluded), 13 (5.1%) had at least one false positive peak. Four of these reactions contained UPTC DNA and were positive for Csubantarcticus as well as Campylobacter and ClariUPTC. *C. subantarcticus* and UPTC belong to the *C. lari* group (Miller et al. 2014) so some similarity of gene sequence could be expected. BLAST searches demonstrated that there was no significant similarity between the Csubantarcticus probe and the genomes of UPTC or the Enterobacteria phage M13 used to extend the lengths of the RPOs. Testing UPTC DNA against the Csubantarcticus and ClariUPTC probes separately would establish whether the observed product was due to an interaction between the UPTC DNA and the Csubantarcticus probe or an artefact associated with the ClariUPTC probe. Since the ClariUPTC and Csubantarcticus probes were both in the Epsilo B probemix, it was not possible to undertake this investigation within this project. It is likely that either the Csubantarcticus or ClariUPTC probe will need to be replaced to produce unambiguous results from unknown samples. Until that happens, any sample producing peaks consistent with both ClariUPTC and Csubantarcticus should be treated with caution.

Extra peaks, corresponding to the next smaller probe, were sometimes observed with control DNA using the Epsilonproteobacteria MLPA assay. This included the Arcobacter peak for

C. subantarcticus (and UPTC) DNA and the Cculiculorum peak for *C. coli* DNA. These extra peaks were also observed when the *C. subantarcticus* DNA was tested with just the A099 Epsilo B probemixes separately even though this probemix does not contain the Arcobacter probe. This phenomenon, with extra peaks 10-20 nucleotides shorter than expected peaks, has been observed by MRC-Holland and may result from folding of the MLPA product or problems with the gel (Paul van Vught, 2016, personal communication). The majority of these extra peaks were excluded from being assigned following optimisation of the position tolerance setting in the band matching algorithm. Some “aberrant” peaks were still observed even once the setting was optimised so care must be taking in interpreting samples positive for two consecutive peaks. Future work would involve changing the length of the probes, possibility by using different M13 vectors, to ensure all “aberrant” peaks can be distinguished from true peaks.

An additional four MLPA reactions had false positive peaks. The 2 ng/µL concentration of DNA from Lasto104.93, a *C. concisus* GS2 strain, was positive for CconcisusGS1. Similarly, the 2 ng/µL concentration of DNA from Lasto24.99, a *C. concisus* GS1 strain, was positive for CconcisusGS2. The 2 ng/µL concentration of *H. fennelliae* CCUG 18820^{T} was positive for Arcobacter. The DNA extract for the *Clostridium difficile* NZRM 2390^{T} was positive for Cculiculorum. Without an alternative explanation for these false positives, it seems likely they were the result of cross contamination introduced during one of the six steps in the MLPA procedure where the tubes are opened. MRC-Holland recommend including at least five probes in a probemix (MRC-Holland 2015, p 17). The target audience for MRC-Holland is human genetic disease researchers, so this recommendation could have been included to ensure the robust detection of possible anomalies. It is also possible the recommendation was included because amplification artefacts, indistinguishable from true positive peaks, had been observed for probemixes containing only a few probes. To address the possibility of low-probe amplification artefacts, a series of positive control probes could be designed that ensure each sample, even those negative for target taxa, have probes amplified.

The detection limit of the Epsilonproteobacteria MLPA may be as high as 1 ng per reaction although freshly prepared dilutions of the DNA were consistently detected at 100 pg per reaction. This suggests the detection limit of freshly prepared dilutions may be lower. The weight of 1000 bp of DNA has been estimated to weigh $1x10^{-18}$ g (Lehninger 1982). The genomes for the strains used to evaluate the detection limit had genomes of between 1.7 Mbp (*C. jejuni* and *C. avium*) and 2.1 Mbp (*C. concisus*). The higher detection limit (1 ng) represents $4.8x10^5$ (*C. concisus*) and $5.9x10^5$ (*C. jejuni* and *C. avium*) copies of the genome. It is, however, likely to be significantly higher than the $6.6x10^{-7}$ µg/mL (600 ag/µL) detection limit reported for a *Campylobacter* real-time PCR (Fachmann et al. 2015) but is comparable to the 0.5-5 pg per reaction and 50-500 pg per reaction detection limits of other reported MLPA assays (Chung et al. 2012, Kim et al. 2016). The poorer detection limit observed for dilutions stored for several days before use highlights the need to store DNA extracts in a more concentrated form with fresh dilutions prepared just before use if the concentrations are approaching the detection limit of an assay.

The Epsilonproteobacteria MLPA assay has the potential to detect taxa in three Epsilonproteobacteria genera in a manner that is suitable for the routine testing of human stool samples. Two PCR-RFLP methods have been published that provide species-level identification for species in all three genera (Gonzalez et al. 2006, Marshall et al. 1999) but the PCR-RFLP format is not suitable for mixed populations. PCR-denaturing gradient gel electrophoresis method has the ability to detect taxa in all three genera (Cornelius et al. 2012, Petersen et al. 2007) but the method is not suitable for routine analysis. Twenty four of the 28 Epsilonproteobacteria MLPA probes produced concordant results when EDTA was limited, an optimal position tolerance was used for band matching and “aberrant” peaks were ignored. The Clariconcheus and Hpullorum probes need to be replaced as they failed to detect DNA from their target taxa. The Cupsaliensis probe either needs to be replaced or an additional probe designed to detect the *C. upsaliensis* strains that do not have the *hae*IIM gene, and further investigation is required to establish whether the ClariUPTC and/or Csubantarcticus probes also need to be

replaced. Although not all of the probes provided sensitive, specific and repeatable results, and further optimisation is required, the MLPA assay shows promise for the simultaneous detection of a range of Epsilonproteobacteria taxa.

4.6 Contributions

Angela Cornelius generated the local BLAST databases and the list of target taxa. She designed the `listAandB.pl` and `centroidRetrieve.pl` scripts written by Associate Professor Patrick Biggs and established the default settings for `listAandB.pl`. Angela also performed all of the LS-BSR, compare_bsr, listAandB and centroidRetrieve analyses; BLAST searches that identified taxon-specific CDS. All of the MLPA probe and probemix design was performed by Angela and she added the DNA to the analysis trays for most of the MLPA runs. The BioNumerics analysis was checked by Angela and she performed the band matching and ran the custom script used to generate the ROC. She also generated all of the tables and figures except Figure 21 and interpreted all of the results in this chapter.

Associate Professor Patrick Biggs wrote the `listAandB.pl` and `retrieveCentroids.pl` scripts.

Bacterial strains were kindly provided by Professor Olivier Vandenberg and Dr Jacqueline Keenan.

DNA was kindly provided by Dr William Miller, Dr Francis Megraud, Dr Maria Figueras, Professor Olivier Vandenberg, Dr Anne Midwinter and Dr Krunoslav Bojanic.

Whole genome sequences were kindly provided before publication by Dr Krunoslav Bojanic, Dr Mohsina Huq and Dr William Miller.

Ms Maud Fazzari resuspended most of the collaborators control DNA that was sent dried on filter paper and performed the majority of the Qubit analyses.

Mrs Susan Lin prepared cultures and performed the DNA extractions to generate some of the control DNA and performed the majority of the MLPA and PCR analyses. She also imported

the majority of the MLPA results into BioNumerics and performed the initial quality checks. Many of the Nanodrop and Qubit analyses on the control DNA were also performed by Susan.

Ms Marilyn Piercy performed the capillary electrophoresis analysis of MLPA products and the 16S rRNA sequencing of control DNA.

Dr Darren Smalley also prepared cultures and performed the DNA extractions to generate some of the control DNA.

Fatemeh Ashari Ghomi performed the pfam searches.

Dr Nicole Wheeler performed the DBS analysis.

Dr David Wood wrote the R script used to generate the receiver operator characteristic from an Excel workbook containing multiple worksheets.

Chapter 5 Childhood Gastroenteritis Case Control Study

5.1 Abstract

Each year billions of children experience acute gastroenteritis globally, resulting in over a million deaths. Current laboratory methods fail to find an aetiological agent in a significant proportion of faecal samples. The bacterial class Epsilonproteobacteria contains several species known to cause gastroenteritis but the role many more species play in this illness is unclear. To better understand this relationship, a Multiplex Ligation-dependent Probe Amplification (MLPA) assay for the detection of 28 taxa within the Epsilonproteobacteria was applied to a collection of DNA extracts from a Belgian childhood gastroenteritis case control study. Using culture-based methods, *Campylobacter jejuni* was isolated from 26 (14.1%) cases and 3 (1.7%) controls, while *C. concisus* was isolated from 6 (3.3%) cases and 4 (2.3%) controls. In contrast, *C. concisus* was the most common species detected by MLPA in both cases (23; 13.2%) and controls (14; 8.6%) and *C. jejuni* was detected in 17 (9.8%) cases and 1 (0.6%) control. The prevalence of *C. jejuni* was significantly higher in cases than controls after adjusting for confounding by age and the presence of *C. concisus* GS2 using culture (p = 0.012, Odds Ratio [OR] = 14.3), MLPA (p = 0.007, OR = 8.3) and a composite gold standard (p = 0.005, OR = 9.0). The difference in prevalence of *C. concisus* was not significant by either culture or MLPA. When individual genomospecies were considered, *C. concisus* genomospecies 2 (GS2) prevalence was significantly higher in cases than controls after adjusting for confounding by age and the presence of *C. jejuni* by MLPA (p = 0.038, OR = 4.0) but not by culture (p = 0.060, OR = 3.5) nor composite gold standard (p = 0.063, OR = 3.5). Fifty-four faecal extracts (16.0%) had discordant results between culture and MLPA, and one sample for which no faecal extract was available for MLPA analysis, had discordant isolate MLPA and PCR results. The PCR supported the MLPA result for the majority (38/55; 69.1%) of

discordant samples. The MLPA assay had an estimated sensitivity of 74.6%, a specificity of 97.8%, a positive predictive value of 89.3%, and a negative predictive value of 93.9% relative to a composite gold standard. Equivalent values for culture are not possible because all culture-positive samples were considered positive in the composite gold standard. These results support *C. jejuni* as the most common Epsilonproteobacterial cause of childhood gastroenteritis and provides some evidence that *C. concisus* GS2 may also be involved in the aetiology of gastroenteritis in children in Belgium.

5.2 Introduction

Gastroenteritis is defined as inflammation of the stomach and intestine (Anonymous 1949) and is often the result of bacterial, viral or parasitic infection. Symptoms typically include diarrhoea, stomach cramps, nausea and vomiting (Adlam et al. 2011). An estimated 2.39 billion cases of diarrhoeal disease occurred globally in 2015 (GBD 2015 Disease and Injury Incidence and Prevalence Collaborators 2016), and approximately 4.66 million episodes occur each year in New Zealand (Adlam et al. 2011). The highest prevalence of acute gastroenteritis is observed for children under five years with an estimated 1.5 billion cases and 1.5-2.5 million deaths globally each year (King et al. 2003). This age group also has the highest prevalence of acute gastroenteritis in New Zealand (Adlam et al. 2011).

Campylobacter jejuni and *C. coli* are collectively the most commonly reported cause of bacterial gastroenteritis worldwide (Anonymous 2012) and account for over a third of all notifications of enteric disease in New Zealand (Health Intelligence Team 2014, 2015, 2016, 2017). The rate of campylobacteriosis in New Zealand during 2012 (158.3 per 100,000 inhabitants) (Health Intelligence Team 2014) was over twice the rate observed in Belgium (68.28 cases per 100,000 inhabitants as calculated from 7,598 cases for a population of 11.13

million[56]) for the same year (Maertens de Noordhout et al. 2017) and the New Zealand rate for 2016 was almost unchanged (158.9 cases per 100,000 inhabitants) (Health Intelligence Team 2017).

C. fetus subsp. *fetus* and *C. upsaliensis* have also been shown to cause human gastroenteritis (Lastovica and Skirrow 2000) and number of other *Campylobacter* taxa, including *C. concisus, C. curvus, C. hyointestinalis, C. jejuni* subsp. *doylei, C. lari* subsp. *lari, C. rectus, C. sputorum* and *C. ureolyticus* have been associated with human gastroenteritis but a causal role has yet to be established (On 2013). Other taxa in the Epsilonproteobacterial class with an established or suspected role in human gastroenteritis include *A. butzleri, A. cryaerophilus* (Vandamme et al. 2005c), *A. skirrowii* (Wybo et al. 2004), *H. bilis, H. canadensis, H. canis, H. cinaedi, H. fennelliae* (On et al. 2005), and *H. pullorum* (Melito et al. 2000).

C. concisus and *C. ureolyticus* have been detected or isolated from both healthy and diarrheic patients (Collado et al. 2013, Cornelius et al. 2012, Engberg et al. 2000, Van Etterijck et al. 1996) and have also been shown to be genetically heterogeneous (Bullman et al. 2013, Vandamme et al. 1989) leading to the hypothesis that some strains have greater potential to cause illness (Aabenhus et al. 2005, Bullman et al. 2012, Deshpande et al. 2013, Kalischuk and Inglis 2011).

Koch's postulates, published in 1891, stipulate that in order to infer an agent causes a disease it must be present in every case of the disease, it must be specific for the disease, and once isolated in pure culture it can reproduce the disease in a naïve host (Firth and Lipkin 2013, Rivers 1937). A more modern approach to inferring causation involves several levels of confidence in a causal relationship (Lipkin 2010). The first level, possible causal relationship, involves the statistical association between an agent and a disease (Lipkin 2010). Case control studies investigate the relationships between potential risk factors and the disease (Silman and

[56] Obtained from the following website on 3rd October 2017 https://www.google.co.nz/publicdata/explore?ds=d5bncppjof8f9_&met_y=sp_pop_totl&idim=country:BEL:NLD:CHE&hl=en&dl=en

Macfarlane 2002, 37) and can help to distinguish between pathogenic and non-pathogenic microorganisms on the premise that unassociated organisms should be equally present in diseased and healthy individuals (Firth and Lipkin 2013).

Reported rates of pathogen (including bacteria, viruses and parasites) detection from stool samples globally varies from 6.4% (de Boer et al. 2010) to 98% (Friesema et al. 2012), although these more extreme values should be considered as outliers as most reported rates were between 20% and 75% (Amar et al. 2007, Boga et al. 2004, Bresee et al. 2012, Cheun et al. 2010, Colomba et al. 2006, Coupland et al. 2013, de Wit, Koopmans, Kortbeek, van Leeuwen, Vinje, et al. 2001, de Wit, Koopmans, Kortbeek, Wannet, et al. 2001, Fiedoruk et al. 2015, Klein et al. 2006, Lake et al. 2009, Lausch et al. 2017, Lorrot et al. 2011, McAuliffe et al. 2013, Olesen et al. 2005, Tam et al. 2012, Tompkins et al. 1999). Although reasons for the differences are not always clear, likely sample-associated influences include the age of the patient, the delay since the onset of symptoms, the severity of the illness, the amount of sample provided and the time and storage conditions between collection and testing. Laboratory-associated variables that could affect the pathogen detection rate include the reason for testing. For example, a relatively small number of tests may be conducted regularly, especially from community-submitted samples, but far more comprehensive testing is likely to be conducted during epidemiological studies or for research.

Standard methods employed in most clinical laboratories worldwide for the isolation of members of the Epsilonproteobacteria limit the taxa detected to *C. jejuni* and *C. coli* (Lindblom et al. 1995, Parker et al. 2006). Maximum recovery of clinically important Epsilonproteobacteria is achieved by including both non-selective filtration and selective methods (Lopez et al. 1998, Van Etterijck et al. 1996, Vandenberg et al. 2004).

Rapid and cost effective detection methods will facilitate a better understand the role of emerging species in human gastroenteritis. Molecular methods offer the potential to detect a broad range of bacterial species in a manner that is not biased by culture conditions. A variety of methods have been published for the molecular detection of the established or potential

human pathogenic Epsilonproteobacterial species in human stool samples but they generally detect either a single taxon, fewer than 10 taxa per reaction or are time-consuming and not suitable for routine use. Multiplex ligation-dependent probe amplification (MLPA), (Figure 6) is a modification of PCR that allows up to 40 genes to be targeted within a single reaction (Schouten et al. 2002).

As a first step in testing whether there are causal relationships between some of the Epsilonproteobacterial taxa and childhood gastroenteritis, the MLPA assay developed for the detection of 28 Epsilonproteobacterial taxa (Chapter 4: Multiplex Ligation-dependent Probe Amplification) was used as an additional diagnostic tool for a Belgian childhood gastroenteritis case control study. In this prospective case control study the outcome of interest is the presence or absence of acute gastroenteritis and the risk factors of interest are the presence or absence of various Epsilonproteobacterial taxa.

5.3 Materials and Methods

5.3.1 Study Design and Identification of Cases and Controls

A gastroenteritis case control study was undertaken in the Belgian city of Brussels which has a population of 1.2 million. Children under 16 years of age presenting at the paediatric emergency rooms (ER) at St. Pierre University Hospital and Queen Fabiola Children's University Hospital with acute gastroenteritis (AG) were recruited prospectively from May 2015 to October 2016. The first 15 cases per week were included in the study. AG was defined as a decrease in stool consistency (loose or liquid) and/or increase in stool frequency (typically at least three per 24 h), with or without fever or vomiting, and lasting fewer than seven days (Guarino et al. 2014). Controls matched to cases by age (within two months) were selected from children attending general paediatric clinics at the same hospital in the following week. Current antibiotic treatment was the only exclusion criterion. Care-givers were asked for consent and,

for controls, suitable containers were posted to the participants to facilitate sample submission at the time of their consultation.

5.3.2 Questionnaires

Questionnaires were completed for each participant. The questionnaires for cases were completed either during the ER visit or later via telephone. The questionnaire was administered to controls via telephone. Follow-up telephone calls were made a month later to fill in missing data and establish the total duration of the AG episode for cases and to establish whether controls had experienced AG after the stool sampling.

The same questionnaire was used for both cases and controls and included questions about the patient's medical history (chronic illnesses, rotavirus vaccination, medications, and antibiotic treatment during the last six months), the AG episode (number of stools per day, presence of blood in the stool, fever, nausea/vomiting, abdominal pain, treatment received, and the necessity of hospitalisation) and risk factors (recent travel, school or day-care attendance, presence of other children at home, pets at home, and diet). A small subset of the variables recorded on the questionnaire was considered in this study to limit the impact of missing data evident for some variables, and in order to focus on testing specific hypotheses related to the presence or absence of pathogens, whilst adjusting for potential confounding variables.

5.3.3 Laboratory Analysis

Routine microbiological methods, including macroscopic and microscopic examination, culture and immunoassays, were used to detect a broad range of bacterial, viral and eukaryotic pathogens associated with gastroenteritis (Tilmanne et al. 2018). Two specific culture protocols were used for the isolation of *Campylobacter* spp. and related organisms. Butzler's medium (Thermofisher scientific, Erembodegem, Belgium) was incubated at 42°C in a microaerophilic atmosphere (10% CO_2, 5% O_2, 0% H_2, 85% N_2). Secondly, a filtration method was performed on Columbia agar containing 5% sheep blood (CBA, Becton Dickinson, Erembodegem,

Belgium) using 0.45 µm pore size cellulose acetate filters (Porafil, Duren, Germany) or 0.6 µm Nuclepore polycarbonate filters (Whatman, Overijse, Belgium) and a microaerophilic atmosphere (10% CO_2, 5% O_2, 0% H_2 and 85% N_2) at 37°C. Colonies typical of *Campylobacter* or related organisms were identified by MALDI-TOF Mass Spectrometry using the Microflex LT and the IVD 2.2 Biotyper database (Bruker, Bremen, Germany) (Martiny et al. 2011).

DNA was extracted from stool samples and isolates using the QiaSymphony (Qiagen, Hilden, Germany) with the elution buffer AVE (RNase-free water with 0.04% sodium azide). Extractions were performed weekly with a maximum time of 10 days between sampling date and extraction of the DNA. The extracted DNA was stored at -20°C and transported frozen to New Zealand where frozen storage was continued until ready for analysis. Each DNA extract was diluted 1 in 100 using 10 mM Tris (pH 8.0) within 24 h of testing and both the undiluted and diluted solutions were analysed using the Epsilonproteobacteria MLPA. The MLPA procedure was the same as described in Chapter 4: Multiplex Ligation-dependent Probe Amplification.

DNA extracts from samples that were positive for an Epsilonproteobacterial species by culture and/or MLPA were tested by one or more of 10 published taxon-specific PCR. Details of the assays are summarised in Table 13. The *C. jejuni/C. coli* multiplex PCR that detects the *C. jejuni lpx*A gene, the *C. coli ceu*E gene and the 23S rRNA gene of thermophilic *Campylobacter* was undertaken as described by Wong (2004) and had previously been validated for use in our laboratory and verified in a study that compared 31 PCR assays (On, Brandt, et al. 2013). DNA from the target taxa and a range of phylogenetically related non-target taxa were included for the remaining nine PCR to provide some information about the taxonomic specificity of the assays. Six annealing temperatures, from 55°C to 65°C in 2°C increments, were used for the *C. cuniculorum* PCR as this information was omitted in the publication of the PCR. Amplicons were detected using the MCE-202 MultiNA microchip electrophoresis system (Shimadzu Corporation, Kyoto, Japan) using standard operating procedures for on-chip mixing and the DNA 1000 reagent kit (Shimadzu).

Table 13: Summary of the Taxon-specific PCR used to Test Some DNA Extracts

Target Taxa	Target Genes	Primer Names	Reaction chemistry	Thermal profile	Product (nt)	Reference
A. butzleri	*qhn*DH	ddAbutzF + ddAbutzR	1 X PCR Buffer II (50 mM KCl, 10 mM Tris, pH 8.3), 250 µM each dNTP, 2.5 mM $MgCl_2$, 200 nM each primer, 1.25 U AmpliTaq Gold	Basic PCR[a] (65°C)	137	(Webb et al. 2016)
C. concisus	*hsp*60	JH0023 + JH0024		Basic PCR (64.6°C)	158	(Chaban et al. 2009)
C. cuniculorum	*rpo*B	CUNrpoB1F + CUNrpoB2R		Basic PCR (55-65°C)	200	(Revez et al. 2013)
C. helveticus	16S rRNA	CHCU146F + CH1371R		Basic PCR (60°C)	1225 or 1375	(Linton, Owen, and Stanley 1996)
C. lari	*hsp*60	JH0015 + JH0016		Basic PCR (54.3°C)	180	(Chaban et al. 2009)
C. upsaliensis	*hsp*60	JH0019 + JH0020		Basic PCR (65°C)	110	(Chaban et al. 2009)
C. ureolyticus	*hsp*60	CureoHSP60F + CureoHSP60R		Basic PCR (58°C)	429	(Bullman et al. 2011, Koziel et al. 2012)
Campylobacter	16S rRNA	C412F + C1228R		Basic PCR (55°C)	816	(Linton, Owen, and Stanley 1996)[b]
C. concisus GS1	23S rRNA	MUC1 + CON1	1 X PCR Buffer II (50 mM KCl, 10 mM Tris, pH 8.3), 250 µM each dNTP, 2.5 mM $MgCl_2$, 0.2 mg/mL BSA, 200 nM each primer, 1.25 U AmpliTaq	94°C for 5 min; 30 cycles of 94°C for 1 min, 60°C for 1 min, 72°C for 1 min; 72°C for 5 min	306	(On, Siemer, et al. 2013)
C. concisus GS2	23S rRNA	MUC1 + CON2			308	

Target Taxa	Target Genes	Primer Names	Reaction chemistry	Thermal profile	Product (nt)	Reference
C. coli	*ceu*E	CeuE forward + CeuE reverse	1 X PCR Buffer II (50 mM KCl, 10 mM Tris, pH 8.3), 250 µM each dNTP, 4 mM $MgCl_2$, 0.2 mg/mL BSA, 400 nM each of ceuE primers, 100 nM each of lpxA and 23S rRNA primers, 1.25 U AmpliTaq	94°C for 3 min; 40 cycles of 94°C for 1 min, 60°C for 1 min, 74°C for 1 min; 74°C for 8 min	695	(Wong et al. 2004)
C. jejuni	*lpx*A	LpxA forward + LpxA reverse			99	
Thermophilic *Campylobacter*	23S rRNA	Therm 1M forward + Therm 2M reverse			246	

[a] 95°C for 10 min; 30 cycles of 95°C for 30 sec, annealing temperature for 1 min, 72°C for 1 min; 72°C for 10 min

[b] Inglis and Kalischuk (2003) reported a typographical error in the paper by Linton, Owen, and Stanley (1996): primer C1288R should read C1228R

5.3.4 Ethical Approval

Approval was obtained for the project entitled “Detection of emerging campylobacter in children with gastroenteritis” (B076201524271), and the associated modification regarding analysis of part of the sample overseas, from the Ethics Committee of the Centre Hospitalier Universitaire Saint-Pierre (Licence No O.M.007) on 8th April 2015.

The analysis of DNA extracts from the stool samples and isolates associated with the case control study by ESR was lodged with the Massey University Human Ethics Committee in August 2015. This project has been evaluated by peer review and judged to be low risk. Consequently, it has not been reviewed by one of the University’s Human Ethics Committee. Angela Cornelius was responsible for the ethical conduct of this research.

5.3.5 Statistical Analysis

Composite gold standard results were created in Excel (2013; Microsoft, Redwood, WA) whereby samples were scored positive for a species if positive by either culture or MLPA and taxon-specific PCR. Test performance characteristics were calculated using the MEDCALC online calculator[57].

In order to account for the pairwise matching of cases and controls, conditional logistic regression analysis was conducted, considering the Epsilonproteobacterial results as the risk factors of primary interest, and adjusting for residual confounding by age. All analyses were carried out using the survival package[58] in R 3.4.1 (2017-06-30) - "Single Candle" (R Core Team 2017).

[57] https://www.medcalc.org/calc/diagnostic_test.php

[58] https://github.com/therneau/survival

5.4 Results

5.4.1 Study Participants

A total of 361 children participated in this study. Of these, 184 (51%) were cases and 177 (49%) were controls. The participants ranged in age from 21 days to 11 years old. The average age for both cases and controls was 2.67 years with 101 (54.9%) of the cases and 96 (54.2%) of the controls being males.

Epsilonproteobacteria culture was undertaken on all 184 cases and 175 of the controls. DNA extraction was not possible for a small number (24, 6.6%) of the stool samples due to insufficient sample or the accidental disposal of the sample before extraction was undertaken. The DNA extracts available for Epsilonproteobacteria MLPA were from 174 cases and 163 controls with average ages of 2.70 and 2.71 years, respectively. As for the cultured samples, a slightly higher proportion of the participants were males with 96 (52.9%) and 92 (56.4%) extracts from male cases and controls, respectively.

5.4.2 Laboratory Results

Cellulose acetate filters with 0.45 µm pore size were used for the first 13 months of the study and they were then replaced with Nuclepore polycarbonate filters which have a pore size of 0.6 µm for the last four months of the study. This decision was outside the control of the culture-based and MLPA-based study leaders. The difference in proportion of *C. concisus*-positive samples for the two filters was not statistically significant using the z-test (data not shown).

The Epsilonproteobacteria culture and MLPA results are summarised in Table 14. *C. jejuni* was isolated from 26 cases and 3 controls, making it the most commonly isolated Epsilonproteobacterial species in this study. Other Epsilonproteobacterial species recovered included *C. concisus* from six cases and four controls, and *A. butzleri* and *C. coli* which were isolated from one control each.

C. concisus was the most commonly detected Epsilonproteobacterial species using MLPA, followed by *C. jejuni*. *C. concisus* GS1 was detected in stool samples from one case and one control that were also positive for *C. concisus* GS2, with the case also being positive for *Salmonella* by culture. *C. coli, C. ureolyticus* and *Campylobacter* sp. were each detected in stool samples from one control. *C. concisus* genomospecies 1 or 2 were detected for over half (21/37, 56.8%) of the samples positive for *C. concisus* by MLPA.

Rotavirus was detected in samples from 20 of the 182 (10.9%) cases and none of the 178 controls tested. One of these rotavirus-positive samples was positive for *C. concisus* by culture and *C. concisus* GS2 by MLPA. Two other rotavirus-positive samples were positive for *C. concisus* by MLPA and a third was positive for *C. concisus* GS2.

Table 14: Summary of Epsilonproteobacteria Culture, Multiplex Ligation-dependent Probe Amplification (MLPA) and Composite Gold Standard Results for the Belgian Childhood Gastroenteritis Case Control Study

	Culture		MLPA		Composite Gold Standard	
	Cases (n = 184)	Controls (n = 175)	Cases (n = 174)	Controls (n = 163)	Cases (n = 174)	Controls (n = 161)
A. butzleri	0	1 (0.6%)	0	0	0	1 (0.6%)
C. coli	0	1 (0.6%)	0	1 (0.6%)	0	1 (0.6%)
C. concisus	6 (3.3%)	4 (2.3%)	23 (13%)	14 (8.6%)	22 (12.6%)	14 (8.7%)
C. concisus GS1			1 (0.6%)*	1 (0.6%)*	0	0
C. concisus GS2			15 (8.6%)	6 (3.7%)	13 (7.5%)	5 (3.1%)
C. jejuni	26 (14.1%)	3 (1.7%)	17 (9.8%)	1 (0.6%)	25 (14.4%)	3 (1.9%)
C. ureolyticus			2 (1.1%)†	1 (0.6%)	0	1 (0.6%)
Campylobacter sp.			0	1 (0.6%)	0	1 (0.6%)
No Epsilonproteobacteria	152 (82.6%)	166 (94.9%)	137 (78.7%)	146 (89.6%)	127 (73.0%)	141 (87.6%)

* both samples were also positive for *C. concisus* GS2 by MLPA and the case was also positive for *Salmonella* by culture,

† both samples were also positive for *C. concisus* GS2 by MLPA and *Salmonella* by culture

5.4.3 Taxon-Specific PCR

Published taxon-specific PCR were used to provide an independent molecular detection method that could be used as part of a composite gold standard to assess the performance of the Epsilonproteobacteria MLPA assay. The taxonomic specificity of nine of the PCR was first assessed using DNA from type or reference strains from the target taxon, and related species, because these PCRs had not previously been used in our laboratory. The taxonomic specificity results for the taxon-specific PCR are summarised in Table 15. Six of the nine PCRs had the expected specificity. The GS1 primer pair for the *C. concisus* GS1 & GS2 PCR, which detects the 23S rRNA gene, produced a product of the expected length for *C. showae* DNA in addition to DNA from *C. concisus* GS1. The *C. lari* PCR, which detects the *hsp*60 gene, produced a product of the expected length for DNA from *C. insulaenigrae* and *C. subantarcticus* in addition to DNA from *C. lari*.

These unexpected results for taxa other than the target were investigated by aligning the primer sequences with the nucleotide sequences of the target genes in the non-target taxa. Figure 26 illustrates the similarity between the MUC1 and CON1 primers used to detect *C. concisus* GS1 strains and the 23S rRNA genes from *C. showae* strains. Figure 27 illustrates the similarity between the JH0015 and JH0016 primers used to detect the *C. lari* and the *hsp*60 genes of *C. insulaenigrae* and *C. subantarcticus*.

Table 15: Taxonomic Specificity Results for Nine Taxon-specific PCR

Strain	Taxon	*A. butzleri* PCR	*C. concisus* PCR	*C. concisus* GS1 & GS2 PCR	*C. cuniculorum* PCR	*C. helveticus* PCR	*C. lari* PCR	*C. upsaliensis* PCR	*C. ureolyticus* PCR	*Campylobacter* genus PCR
RM 15224^T	*A. bivalviorum*	-								
CCUG 30485^T	*A. butzleri*	+								
CECT 7203^T	*A. cibarius*	-								
RM 15227^T	*A. cloacae*	-								
CCUG 17801^T	*A, cryaerophilus*	-								
RM 14018^T	*A. defluvii*	-								
RM 15222^T	*A. ellisii*	-								
RM 5350^T	*A. halophilus*	-								
RM 14021^T	*A. marinus*	-								
RM 14015^T	*A. molluscorum*	-								
RM 14013^T	*A. mytili*	-								
RM 3221^T	*A. nitrofigilis*	-								
CCUG 10374^T	*A. skirrowii*	-								
RM 15228^T	*A. suis*	-								
RM 5348^T	*A. thereius*	-								
RM 12658^T	*A. trophiarum*	-								
RM 16046^T	*A. venerupis*	-								
RM 8639^T	*C. avium*				-	-	-	-		+
CCUG 54429^T	*C. canadensis*									+
CCUG 11283^T	*C. coli*									+
CCUG 13144^T	*C. concisus* GS1		+	1						+
CCUG 19995	*C. concisus* GS2		+	2						
CCUG 56289^T	*C. cuniculorum*				+	-		-		+
CCUG 13146^T	*C. curvus*		-	-						
CCUG 30682^T	*C. helveticus*				-	+		-		+
CH001A^T	*C. hominis*		-							
CCUG 14169^T	*C. hyointestinalis*									+
LMG 22716^T	*C. insulaenigrae*						+			+
CCUG 11284^T	*C. jejuni*						-			+
CCUG 23947^T	*C. lari*						+			+
R-13342^T	*C. peloridis*						-			+
CCUG 30254^T	*C. showae*		-	1						
CCUG 38513^T	*C. subantarctius*						+			+
CCUG 9728^T	*C. sputorum*								-	
NZRM 3675^T	*C. upsaliensis*				-	-	-	+		+
CCUG 7319^T	*C. ureolyticus*								+	+
RM 9726^T	*C. volucris*									+

^{T} type strain, + positive, - negative, 1 GS1, 2 GS2, 1+2 GS1 + GS2

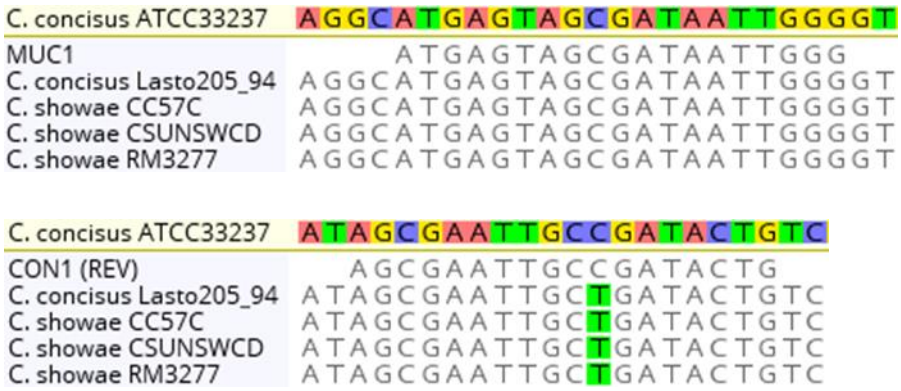


Figure 26: Geneious Alignment of the MUC1 and CON1 Primers and the 23S rRNA Gene Sequences Extracted from Two *C. concisus* GS1 strains and Three *C. showae* Strains

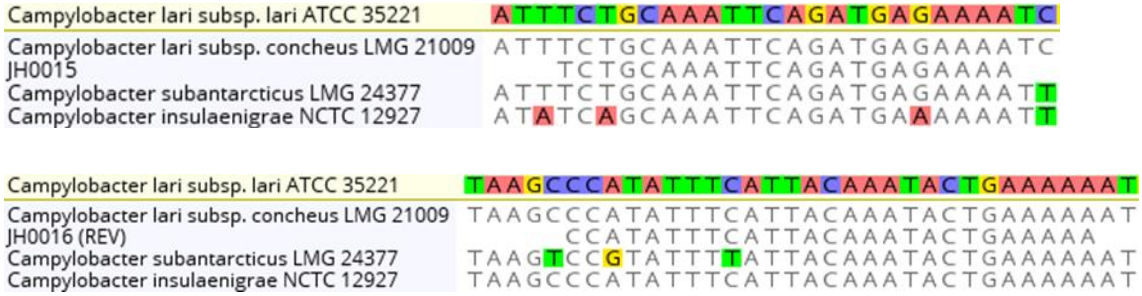


Figure 27: Geneious Alignment of the JH0015 and JH0016 Primers and the *hsp*60 Gene Sequences Extracted from *C. lari* subsp. *lari*, *C. lari* subsp. *concheus*, *C. insulaenigrae* and *C. subantarcticus*

The majority of samples (283/337, 84%) analysed for Epsilonproteobacteria by both culture and MLPA produced concordant results. The DNA extracts from all samples positive for Epsilonproteobacteria by either culture or MLPA were further analysed using at least one of the published taxon-specific PCR summarised in Table 13. The culture, MLPA and taxon-specific PCR results for the 55 discordant samples, including one sample for which no faecal extract was available for analysis by MLPA, are summarised in Table 16. For one sample, *C. concisus* GS2 was detected in the DNA extracted from the faecal sample using both MLPA and PCR but the isolated strain was identified as *C. concisus* GS1 using both MLPA and PCR. Both *C. concisus* and *C. jejuni* were detected by MLPA in a sample from which only *C. jejuni* was isolated. *C. concisus* GS1 and *C. jejuni* were detected using PCR.

Taxon-specific PCR results matched the culture result for 14 (25.5%) of the 55 discordant samples. *C. concisus* was isolated, and detected by PCR, for three samples which were

negative by MLPA. *C. jejuni* was isolated, and detected by PCR, for seven samples that were negative by MLPA. The strains were positive for *C. jejuni* using both MLPA and PCR. The remaining four samples were negative for Epsilonproteobacteria by culture, positive for *C. concisus* by MLPA and negative for *C. concisus* (and GS1 & GS2) by PCR.

Taxon-specific PCR on the DNA extracted from stool samples matched the MLPA results for 38 (69.1%) of the 55 samples with discordant results. *A. butzleri* was isolated from one sample but this was not detected by MLPA or PCR. Similarly, *C. concisus* was isolated from three samples but were negative by MLPA and PCR. *C. jejuni* was isolated from three samples that were negative for this taxa by MLPA and PCR. Two of these samples were negative by MLPA and one was positive for *C. concisus* GS2 which was also detected by PCR. A total of 31 samples (56.4% of the discordant samples) were negative for Epsilonproteobacteria by culture and positive by MLPA and PCR. Twelve of these were positive for *C. concisus* (and no genomospecies) by MLPA. The PCR result for two of the samples was GS1, for three samples was GS1 & GS2 and for seven samples was GS2. Of these 12 samples, only one GS2-positive and two GS1 & GS2 samples were positive by *C. concisus* PCR and only five of eleven samples tested with the *Campylobacter* genus PCR were positive.

Table 16: Summary of Culture, MLPA and Taxon-specific PCR Results for Samples with Discordant Results

Campylobactereaceae by Culture	Faecal MLPA Result	*A. butzleri* PCR	*C. concisus* PCR	*C. concisus* GS1 & GS2 PCR	*C. cuniculorum* PCR	*C. helveticus* PCR	*C. lari* PCR	*C. upsaliensis* PCR	*C. ureolyticus* PCR	*Campylobacter* genus PCR	Multiplex PCR	Isolate MLPA Result	*C. concisus* PCR	*C. concisus* GS1 & GS2 PCR	Multiplex PCR	PCR Matches
C. concisus	*C. concisus* GS2		+	2								*C. concisus* GS1	+	1		N/A
C. jejuni	*C. concisus* + *C. jejuni*		-	1						-	*jej*	*C. jejuni*			*jej*	Both
C. concisus	Negative		-	1+2						+		*C. concisus* GS1	+	1		Culture
*C. concisus**	Negative		-	2						-		*C. concisus* GS2	+	1+2		Culture
*C. concisus**	Negative		-	2						-		*C. concisus*	+	1		Culture
C. jejuni	Negative									+	*jej*	*C. jejuni*			*jej*	Culture
C. jejuni	Negative									-	*jej*	*C. jejuni*			*jej*	Culture
C. jejuni	Negative									+	*jej*	*C. jejuni*			*jej*	Culture
*C. jejuni**	Negative									-	*jej*	*C. jejuni*			*jej*	Culture
C. jejuni	Negative									-	*jej*	*C. jejuni*			*jej*	Culture
C. jejuni	Negative									+	*jej*					Culture
C. jejuni	Negative									-	*jej*	*C. jejuni*			*jej*	Culture
Negative*	*C. concisus*		-	-												Culture
Negative*	*C. concisus*		-	-						-						Culture
Negative*	*C. concisus* GS1 & GS2		-	-						-						Culture
Negative	*C. concisus*		-	-						-						Culture
*A. butzleri**	Negative	-	-	-	-	-	-	-	-	-	-					MLPA
C. concisus	Negative		-	-						-		*C. concisus* GS2	-	2		MLPA
*C. concisus**	Negative		-	-						-		*C. concisus* GS2	+	2		MLPA
*C. concisus**	Negative		-	-						-		*C. concisus* GS1	+	1		MLPA
*C. jejuni**	*C. concisus* GS2		-	2						+	-	*C. jejuni*			*jej*	MLPA
C. jejuni	Negative									-	-	*C. jejuni*			*jej*	MLPA
C. jejuni	Negative									-	-	*C. jejuni*			*jej*	MLPA

Campylobactereaceae by Culture	Faecal MLPA Result	*A. butzleri* PCR	*C. concisus* PCR	*C. concisus* GS1 & GS2 PCR	*C. cuniculorum* PCR	*C. helveticus* PCR	*C. lari* PCR	*C. upsaliensis* PCR	*C. ureolyticus* PCR	*Campylobacter* genus PCR	Multiplex PCR	Isolate MLPA Result	*C. concisus* PCR	*C. concisus* GS1 & GS2 PCR	Multiplex PCR	PCR Matches
Negative	*C. concisus*		-	1						-						MLPA
Negative*	*C. concisus*		-	1						-						MLPA
Negative	*C. concisus*		+	1+2						-						MLPA
Negative*	*C. concisus*		+	1+2						+						MLPA
Negative*	*C. concisus*		-	1+2						-						MLPA
Negative	*C. concisus*		-	2						+						MLPA
Negative	*C. concisus*		-	2						-						MLPA
Negative*	*C. concisus*		+	2						+						MLPA
Negative*	*C. concisus*		-	2						+						MLPA
Negative*	*C. concisus*		-	2												MLPA
Negative	*C. concisus*		-	2						+						MLPA
Negative	*C. concisus*		-	2						-						MLPA
Negative	*C. concisus* GS2		+	1						+						MLPA
Negative	*C. concisus* GS2		-	1+2						-						MLPA
Negative*	*C. concisus* GS2		-	1+2						-						MLPA
Negative	*C. concisus* GS2		-	1+2						+						MLPA
Negative	*C. concisus* GS2		+	1+2						+						MLPA
Negative	*C. concisus* GS2		+	2						+						MLPA
Negative*	*C. concisus* GS2		+	2						+						MLPA
Negative	*C. concisus* GS2		+	2						+						MLPA
Negative	*C. concisus* GS2		-	2						-						MLPA
Negative*	*C. concisus* GS2		+	2						+						MLPA
Negative	*C. concisus* GS2		-	2						+						MLPA
Negative	*C. concisus* GS2		-	2						-						MLPA

Campylobactereaceae by Culture	Faecal MLPA Result	*A. butzleri* PCR	*C. concisus* PCR	*C. concisus* GS1 & GS2 PCR	*C. cuniculorum* PCR	*C. helveticus* PCR	*C. lari* PCR	*C. upsaliensis* PCR	*C. ureolyticus* PCR	*Campylobacter* genus PCR	Multiplex PCR	Isolate MLPA Result	*C. concisus* PCR	*C. concisus* GS1 & GS2 PCR	Multiplex PCR	PCR Matches
Negative*	*C. concisus* GS2		-	2						-						MLPA
Negative	*C. concisus* GS2		+	2						+						MLPA
Negative	*C. concisus* GS2 + *C. ureolyticus*		+	2					-	+						MLPA
Negative	*C. concisus* GS2 + *C. ureolyticus*		+	2					-	+						MLPA
Negative	*C. jejuni*									+	*jej*					MLPA
Negative*	*C. ureolyticus*								+	+						MLPA
Negative*	*Campylobacter* sp.	-	-	-	-	-	-	-	-	+	-					MLPA
C. concisus	Not tested											*C. concisus* GS2	+	1+2		N/A

+ positive, - negative, 1 GS1, 2 GS2, 1+2 GS1 + GS2, *jej C. jejuni*, N/A not applicable, * sample was from a control

C. concisus GS2 was detected by MLPA in 14 samples that were negative for Epsilonproteobacteria by culture and nine of these samples were also positive for this taxa by PCR. Four of the samples were positive for GS1 & GS2 by PCR and one sample was positive for *C. concisus* GS1. Only seven (five GS2-positive and one each positive for GS1 and GS1 & GS2) of these 14 samples were positive by *C. concisus* PCR and nine samples were positive for the *Campylobacter* genus PCR.

Both *C. concisus* GS2 and *C. ureolyticus* were detected in two samples with only GS2 being detected by PCR. Three additional samples were negative for Epsilonproteobacteria by culture and positive for the same taxon by both MLPA and PCR. The taxa detected in these three samples were *C. jejuni*, *C. ureolyticus* and the *Campylobacter* genus.

The isolate MLPA and PCR results were discordant for one sample that did not have a faecal DNA extract tested by MLPA. The MLPA was positive for *C. concisus* GS2 and the PCR was positive for GS1 & GS2.

Composite gold standard results were generated with samples being considered positive if an isolate was recovered by culture or if both MLPA and PCR detected the same species. For *C. concisus* and *C. concisus* GS2, a sample was considered positive for the same taxon detected by MLPA if either the *C. concisus* or GS2 PCR were positive. The GS1 PCR was considered unreliable when the sample was not also positive by *C. concisus* PCR. One sample was positive for *C. jejuni* by culture, *C. concisus* GS2 by MLPA and *C. concisus* by PCR. The composite gold standard for this result was recorded as *C. jejuni* + *C. concisus* GS2 and the MLPA was considered a false negative for the test performance characteristics. The composite gold standard results are summarised in Table 14. The prevalence of *C. jejuni* and *C. concisus* GS2 in cases, based on the composite gold standard, was 14.4% and 7.5%, respectively. Conversely, the prevalence of *C. jejuni* and *C. concisus* GS2 in controls, based on the composite gold standard, was 1.9% and 3.1%, respectively.

5.4.4 Statistical Results

Cases were matched to controls based on age (± 60 days), with the majority (128 of 179, 71.5%) of sets containing one case and one control, 17 sets containing two cases and one control, 10 sets containing one case and two controls, 11 sets containing only one case and 13 sets containing only one control.

The sensitivity of the MLPA, relative to the composite gold standard, for the faecal extracts was 74.6% (95% CI 62.5 – 84.5%) and the specificity was 97.8% (95% CI 95.2 – 99.2%). The positive predictive value for the MLPA was 89.3% (95% CI 78.9 – 94.0%) and the negative predicative value was 93.9% (95% CI 91.1 – 95.9%). Equivalent values for culture are not possible because all culture-positive samples were considered positive in the composite gold standard.

Conditional logistic regression was used to evaluate associations between the Epsilonproteobacterial results and gastroenteritis status (case/control) as the outcome. Table 17 summarises the significant results from the univariate and multivariate analyses. In the univariate analysis each variable was considered individually, the multivariate analysis is a single model including all three variables that gave the largest change in deviance based on the Likelihood Ratio test statistic. Despite matching within 2 month age bands, it was evident that, within each case/control set, the cases were on average younger than the controls.

Any differences in prevalence between cases and controls for the variables gender, *C. concisus* by culture and MLPA and *C. concisus* GS1 (by MLPA) were not statistically significant. Rotavirus was not included in the conditional regression analysis because all of the rotavirus-positive samples were from cases.

Table 17: Summary of the Significant Results from the Conditional Logistic Regression Analysis

Risk Factor	Coefficient (Standard Error)	Odds Ratio	95% Confidence Interval	*p*-value
Univariate analysis				
C. jejuni (culture)	2.45 (0.74)	11.6	2.7 - 49.5	<0.001
C. jejuni (MLPA)	2.63 (1.04)	13.9	1.8 - 106.4	0.011
C. jejuni (CGS*)	2.25 (0.75)	9.5	2.2 - 41.4	0.003
C. concisus GS2 (MLPA)	1.46 (0.64)	4.3	1.2 - 15.1	0.023
C. concisus GS2 (CGS)	1.47 (0.65)	4.3	1.2 - 15.5	0.023
Multivariate analysis 1				
C. jejuni (MLPA)	2.66 (1.06)	14.3	1.8 - 113.1	0.012
C. concisus GS2 (MLPA)	1.39 (0.67)	4.0	1.1 - 14.9	0.038
Age (days)†	-0.02 (0.006)	0.98	0.97 - 0.99	<0.001
Likelihood ratio test = 36.1 on 3 df‡, *p*-value = 7.1×10^{-8}, 27 missing values				
Multivariate analysis 2				
C. jejuni (culture)	2.11 (0.78)	8.3	1.8 - 37.9	0.007
C. concisus GS2 (MLPA)	1.26 (0.67)	3.5	0.95 - 13.0	0.060§
Age (days)	-0.02 (0.006)	0.98	0.97 - 0.99	<0.001
Likelihood ratio test = 35.2 on 3 df, *p*-value = 1.1×10^{-7}, 29 missing values				
Multivariate analysis 3				
C. jejuni (CGS)	2.20 (0.77)	9.0	2.0 - 40.8	0.005
C. concisus GS2 (CGS)	1.26 (0.68)	3.5	0.93 - 13.4	0.063§
Age (days)	-0.02 (0.006)	0.98	0.97 - 0.99	<0.001
Likelihood ratio test = 36.4 on 3 df, *p*-value = 6.3×10^{-8}, 29 missing values				

* composite gold standard; † adjusting for residual confounding given the pairwise matching on age, ‡ degrees of freedom, § not significant (p>0.05)

5.5 Discussion

This study was designed to evaluate a range of Epsilonproteobacterial taxa as risk factors for acute gastroenteritis in children and investigate the suitability and advantages that MLPA could provide in a routine clinical setting. The statistically significant association observed between *C. jejuni* and cases is concordant with this species being well recognised as the most commonly reported cause of bacterial gastroenteritis worldwide (Anonymous 2012). Based on the composite gold standard results for the 174 samples from cases that were analysed by MLPA, the prevalence of *C. jejuni* in stool samples of children with gastroenteritis was 14.4%. This is higher than the combined *C. jejuni/C. coli* prevalence of 4.6% observed in Denmark (Nielsen, Engberg, et al. 2013a) and likely reflects the inclusion of samples from children presenting at general practitioners as well as hospitals in the Danish study (Nielsen, Ejlertsen, et al. 2013).

C. concisus GS2 was the only other Epsilonproteobacterial taxa for which a significant association with cases was observed, although the result was statistically significant in the multivariate analysis that included *C. jejuni* (by MLPA) and not *C. jejuni* (by culture) or *C. jejuni* (by CGS). Multivariate analysis 3, which used results generated using the composite gold standard, was slightly better than the other two. Subtle confounding between the presence of *C. jejuni* and *C. concisus* GS2 and the differences in the number of missing values are likely to be the reasons for the variation. This genomospecies is genetically distinct from the *C. concisus* type strain ATCC 33237 (Aabenhus et al. 2005, Chung et al. 2016, Mahendran et al. 2015, Miller et al. 2012, Nielsen, Nielsen, and Torpdahl 2016, On, Siemer, et al. 2013, Vandamme et al. 1989), as demonstrated in Chapter 4: Multiplex Ligation-dependent Probe Amplification, and includes the well characterised strains CCUG 19995 and ATCC 51561. This is the first case control study investigating the association between genomospecies of *C. concisus* and childhood gastroenteritis. Previous studies have demonstrated a high prevalence of *C. concisus* in children with gastroenteritis (Engberg et al. 2000, Nielsen, Engberg, et al. 2013a,

Vandenberg et al. 2013) and the presence of *C. concisus* in the stool samples of both healthy and diarrhoeic patients (Cornelius et al. 2012, Engberg et al. 2000, Inglis, Boras, and Houde 2011) but genomospecies analysis was not undertaken in these studies. Based on the composite gold standard results for the 174 samples from cases that were analysed by MLPA, the prevalence of *C. concisus* in stool samples of children with gastroenteritis was 13.2%. This is higher than the prevalence of 3.6% observed in Denmark (Nielsen, Engberg, et al. 2013a) and likely reflects the inclusion of samples from children from both the community and hospitals in the Danish study (Nielsen, Ejlertsen, et al. 2013) as well as the use of both culture and molecular (MLPA and PCR) results in the current study. In studies where the clinical manifestations of the two main genomospecies have been evaluated, *C. concisus* GS2 has been associated with immunocompetent patients and/or those without co-infections (Aabenhus et al. 2005), diarrhoeic patients (Kalischuk and Inglis 2011) and inflammatory enteric diseases (Ismail et al. 2012). Isolates of this genomospecies have also been shown to be more enteroinvasive with respect to Caco-2 cells (Ismail et al. 2012) and have higher mean epithelial invasion and translocation (Kalischuk and Inglis 2011) than GS1 isolates. There is also some evidence that *C. concisus* GS2 is more adapted to the human gastrointestinal tract than *C. concisus* GS1 (Wang et al. 2017).

Rotavirus was detected in 20 (10.9%) samples from cases but this data could not be incorporated into the conditional regression analysis because no samples from controls were also positive. The detection of *C. concisus* or *C. concisus* GS2 by culture and/or MLPA in four of these rotavirus-positive samples reinforces the need for caution in associating *C. concisus* GS2 with childhood gastroenteritis.

C. ureolyticus was detected by MLPA in DNA extracts from two (1.1%) cases and one (0.6%) control although only the detection in the control was supported by the *C. ureolyticus* *hsp*60 PCR. This rate of detection is similar to the 1.15% (83/7194) of faecal samples positive for *C. ureolyticus* by *hsp*60 PCR in study involving faecal samples submitted to an Irish hospital from both children and adults (Bullman et al. 2011). Other studies have reported higher rates,

with *C. ureolyticus* being detected by PCR-denaturing gradient gel electrophoresis (PCR-DGGE) in 4.0% (2/50) of faecal samples from non-hospitalised children in Belgium (Vandenberg et al. 2013) and a recent Japanese study detecting *C. ureolyticus* by *hsp*60 PCR in 25.1% (147/586) of rectal swabs from children up to 12 years of age presenting at hospital with diarrhoea (Hatanaka et al. 2017). In addition, *C. ureolyticus* was detected by PCR-DGGE in 10.9% (14/128) of samples from diarrhoeic patients and 25% (12/49) of healthy controls in a small community-based study in New Zealand (Cornelius et al. 2012). All five of the above studies reported the detection of other *Campylobacter* species or known pathogens for at least a proportion of the *C. ureolyticus*-positive samples. In this study both of the cases from which *C. ureolyticus* was detected by MLPA were also positive for *C. concisus* GS2 and *Salmonella,* and both of the *C. ureolyticus*-positive samples in the Belgian study were also positive for *C. hominis* (Vandenberg et al. 2013). In the Irish study 36.1% (30/83) of the *C. ureolyticus*-positive samples were also positive for at least one other *Campylobacter* species (Bullman et al. 2011). Over a third (51/147, 34.7%) of the *C. ureolyticus*-positive samples from Japan were also positive for another enteric pathogen (Hatanaka et al. 2017). One of the 14 (7.1%) *C. ureolyticus*-positive diarrhoeal samples in the New Zealand study was also positive for norovirus, another was positive for *C. jejuni/coli* and eight (57.1%) were positive for other *Campylobacter* species (unpublished data). In addition, other non-*C. jejuni/coli Campylobacter* species were detected in five of the 12 *C. ureolyticus*-positive samples from healthy individuals (unpublished data). Further investigations are required to establish whether *C. ureolyticus* has a causal role in human gastroenteritis.

Two of the taxon-specific PCR used to provide an independent molecular assay for samples positive for Epsilonproteobacteria by either culture or MLPA produced bands of the expected size from non-target taxa. The *C concisus* GS1 PCR, which detects a region of the 23S rRNA gene, was positive for *C. showae* DNA in addition to *C. concisus* GS1 DNA. Amplification of *C. showae* DNA using these primers has previously been reported by Inglis, Boras, and Houde (2011). Alignment of these PCR primers to the 23S rRNA genes extracted

from two *C. concisus* GS1 and three *C. showae* genomes demonstrated that the *C. showae* sequences were indistinguishable from the sequence observed in one of the *C. concisus* GS1 genomes and had only a single nucleotide difference to one of the PCR primers. The *hsp*60 gene from *C. subantarcticus* had 100% identity with the primer JH0015 and two nucleotide differences relative to primer JH0016, including three nucleotides from the 3' end. Similarly, the *hsp*60 gene from *C. insulaenigrae* had the 100% identity with the primer JH0016 and two differences relative to primer JH0015, including three nucleotides from the 3' end. The three nucleotides at the 3' end of a primer are known to be important for primer specificity but there are no hard rules saying these primers wouldn't anneal to the sequences observed in these species. *C. insulaenigrae* was described in 2004 (Foster et al. 2004) but this species was not mentioned in the paper describing these primers designed to detect *C. lari* (Chaban et al. 2009). *C. subantarcticus* was described in 2010 (Debruyne et al. 2010a), after the paper describing the primers. These three species clustered together in the neighbour-joining tree based on partial *hsp*60 gene sequences (Debruyne et al. 2010a). It is likely that the primers designed to detect *C. lari* also detect *C. insulaenigrae* and *C. subantarcticus*. This observation reinforces the recommendation of On, Brandt, et al. (2013) that revalidation of PCR assays be regularly undertaken to ensure they continue to perform to an acceptable degree of accuracy.

There are currently no published molecular assays suitable for the detection of nine taxa included in the Epsilonproteobacteria MLPA assay (*C. avium, C. insulaenigrae, C. jejuni* subsp*. doylei, C. lari* subsp*. concheus,* UPTC*, C. peloridis, C. subantarcticus, C. volucris* and *H. fennelliae*) in complex samples such as stools and the majority of the molecular assays published for the included taxa detect only one taxon. The highest number of taxa from the Epsilonproteobacteria MLPA assay that were included in a single molecular assay suitable for this type of sample was four (Kamei et al. 2016, Klena et al. 2004, Wang et al. 2002, Yamazaki-Matsune et al. 2007) with only the assay reported by Klena et al. (2004) having been applied to complex samples (Kasper et al. 2012).

Molecular methods such as PCR (and including MLPA) are known to suffer from both false positive and false negative results (Inglis, Boras, and Houde 2011). The Epsilonproteobacteria MLPA was designed to detect 28 taxa but the assay failed to detect DNA from pure cultures of *H. pullorum* and *C. lari* subsp. *concheus* and some strains of *C. upsaliensis* (see Chapter 4 Multiplex ligation-dependent probe amplification). In addition, some probes may not provide the specificity expected and some probe lengths may need to be changed to generate an optimal assay. Of the 37 stool samples positive for *C concisus* by MLPA, 23 (62%) were also positive for the *C. concisus* GS1 and/or *C. concisus* GS2. The absence of a GS-specific MLPA result for the remaining 14 (38%) samples may be due to a) as yet uncharacterised sequence variation in the region detected by the CconcisusGS1 and CconcisusGS2 probes, b) the existence of *C. concisus* strains that belong to additional genomospecies, or c) a lack of specificity for the Cconcisus MLPA probe. Three of each of the nine *C. concisus* GS1 and *C. concisus* GS2 strains tested in Chapter 4 Multiplex ligation-dependent probe amplification were also negative for the expected GS-specific probes on at least one occasion (see Appendix IV). These pure culture results suggest that the CconcisusGS1 and CconcisusGS2 MLPA probes may be more sensitive to inhibition than the Cconcisus MLPA probe. The MLPA, however, produced relatively few false positive results relative to the composite gold standard, as demonstrated by the specificity of 98.5%. Conversely, there were higher than desirable numbers of false negative results relative to the composite gold standard, as demonstrated by the sensitivity of 75.4%. The false negative results may reflect the small sample size able to be extracted for MLPA to avoid PCR inhibitors from the faecal sample and also the non-homogeneous distribution of bacterial cells in faecal samples (Inglis, Boras, and Houde 2011). In addition, the MLPA currently does not contain a control for the ligation step so a proportion of the false negative samples may be due to inhibition of the ligase enzyme.

The majority (32/54, 59.3%) of the faecal extracts with discordant results were negative by culture and positive for *C. concisus* by MLPA. Taxon-specific PCR supported the MLPA results for 28 (87.5%) of these samples. A significant contributor to this result is the lack of

hydrogen in the growth atmosphere used for the recovery of *Campylobacter* spp. and related organisms (Vandamme et al. 2005b). The use of cellulose acetate filters with 0.45 μm pores, rather than polycarbonate filters with 0.60-0.65 μm pores for the majority (13 of 17 months) of the study, may also have contributed to *C. concisus* being isolated from fewer samples than detected by MLPA (and PCR) (Nielsen, Engberg, et al. 2013b, Wareing et al. 1998), although the difference in the proportion of *C. concisus*-positive samples between the two filters in this study was not statistically significant.

One *C. concisus* isolate was GS2 by MLPA and GS1 & GS2 by PCR. This isolate was recovered from a faecal sample that was not analysed using MLPA. To date, WGS has not yet been successful for this isolate so additional confirmation of genomospecies, using ANI, GBDP etc, has not yet been possible. Isolates positive for both GS1 and GS2 have previously been reported (Kalischuk and Inglis 2011) but WGS has not yet been reported for any of these isolates. It is possible that these isolates that are positive for both the GS1 and GS2 PCR are either GS1 or GS2 or they may be part of another, as yet uncharacterised, genomospecies.

This study has at least two potential sources of bias. The selection criteria for study participants introduced spectrum and subgroup bias (also called case mix bias) (Mower 1999). The study focused on children due to the high prevalence of gastroenteritis but this also means the results may not be applicable to the general population. Children presenting at the ER are likely to have more severe symptoms than the general population and the associations observed may not exist for milder illness. In addition, the controls, although not experiencing gastroenteritis, may not reflect the general population of healthy children given that they were attending a general paediatric clinic at the hospital. For example, gut microbial dysbiosis has been associated with atopy, asthma and autism-spectrum disorder (Arrieta et al. 2014, Huang and Boushey 2015), and this microbial imbalance may affect the carriage of Epsilonproteobacterial species. The second bias, incorporation bias, relates to the inclusion of MLPA results in the composite gold standard used for the receiver operator characteristic

analysis and is likely to have resulted in elevated sensitivity and specificity scores (Kohn, Carpenter, and Newman 2013).

Although the sensitivity of the Epsilonproteobacteria MLPA assay needs improvement, it provides a much needed method for evaluating the role of emerging Epsilonproteobacterial species in diseases such as gastroenteritis. This single assay simultaneously detects not only *C. jejuni,* which is detected by most of the commercial syndromic assays currently available, but also several additional taxa known or suspected to have a role in the aetiology of this disease. Detection of these additional taxa may help provide an aetiological agent for some of the many stool samples submitted for microbiological analysis that currently have no cause identified. Future expansion of the MLPA assay to include other enteric pathogens, including viruses, parasites and additional bacteria, will offer an exhaustive diagnosis of gastroenteritis. Other future developments of the assay include optimising the sensitivity of the assay, including additional internal controls to monitor inhibition and minimising the turnaround time for results. MLPA is cost-effective and technically simple enough that a single comprehensive assay could be developed for screening of stool samples in outbreak situations and complete microbiological analysis in resource poor countries. The application of the MLPA assay to the Belgian childhood gastroenteritis case control study has demonstrated that MLPA can be applied to the detection of pathogens in human faecal samples. In addition, it provided some evidence of a possible role for *C. concisus* GS2 in the aetiology of childhood gastroenteritis, however this association should be treated with caution as the number of positive samples was low and rotavirus was also detected in two of the *C. concisus* GS2-positive samples from cases. Larger studies involving the detection of *C. concisus* GS2 are required to establish whether the observed association with cases is significant.

5.6 Contributions

Angela Cornelius arranged for the Belgian samples to be sent to ESR, directed all of the PCR and MLPA analyses performed at ESR, added the DNA to the sample trays for most of the

MLPA runs and performed the statistical analyses. All of the tables and figures in this chapter were generated by Angela and she performed all of the data interpretation.

The Belgian childhood gastroenteritis case control study was designed and conducted by Dr Anne Tillmanne and Professor Olivier Vandenberg. Dr Marie Hallin and Dr Delphine Martiny organised the DNA extractions and shipments to New Zealand.

Mrs Susan Lin performed the majority of the MLPA and PCR analyses. She also imported the majority of the MLPA results into BioNumerics and performed the initial quality checks.

Ms Marilyn Piercy performed the capillary electrophoresis analysis of MLPA products.

Chapter 6: General Discussion

6.1 Summary

This thesis aimed to develop and apply new tools to assess the contribution of a range of Epsilonproteobacterial taxa in human gastroenteritis. Whole genome sequence-based *in silico* analyses were used to explore the classification of the genetically heterogeneous species *Campylobacter concisus* in order to inform the development of the tool. Genomic comparisons were then undertaken on the three main genera to identify taxon-specific coding sequence (CDS) which were then used as targets for a multiplex ligation-dependent probe amplification (MLPA) assay for the detection of 28 taxa within the bacterial class Epsilonproteobacteria. The MLPA assay was then applied to faecal DNA extracts from a Belgian childhood gastroenteritis case control study and provided some evidence that *C. concisus* GS2 infection is a risk factor for acute gastroenteritis in children.

6.2 Whole Genome Sequencing and Bioinformatics

Whole genome sequencing (WGS) has provided the opportunity to obtain the complete genetic information about a microorganism, although doing so currently still requires more time and effort than is justified for routine use. Even when draft rather than complete genomes are generated, as has become almost commonplace with short read sequencing becoming more cost-effective, the nucleotide sequence for a significant proportion of the genome can now be elucidated within two days of isolation. With the application of bioinformatics software, this genetic information, in conjunction with publicly available websites, can provide species level identification for many clinically important bacterial species.

Whole genome sequences clearly have huge promise for microbial taxonomic studies. The current taxonomic definition of a species requires DNA–DNA relatedness of approximately 70% for strains within the same species (Stackebrandt et al. 2002, Tindall et al. 2010, Wayne et

al. 1987). Proposed species boundaries, with equivalence to 70% DNA-DNA relatedness, have been generated for both of the bioinformatic tools used in these studies: average nucleotide identity (ANI (Goris et al. 2007)) and genome BLAST distance phylogeny (GBDP (Meier-Kolthoff et al. 2013)). Care needs to be taken, however, when using these tools as the first step of both analyses is the identification of genomic regions showing sequence similarity. It is therefore possible for very distinctly related genomes to produce high ANI or GBDP results based on a small proportion of the genomes. Including a measure that represents the proportion of the genomes that was included in the analysis, such as the alignment fraction proposed by Varghese et al. (2015), may overcome this shortcoming. ANI and GBDP have been included in the latest minimal standards for describing new species within four Epsilonproteobacterial genera (On et al. 2017). The new minimal standards still recommend a range of physical and biochemical tests be undertaken on strains representing the new species, as well as on the type strains of related species, and it must be possible to phenotypically differentiate the new species from strains of other species (On et al. 2017). The ability to describe a new species within the *Arcobacter, Campylobacter, Helicobacter* and *Wolinella* genera without the need to undertake DNA-DNA reassociation analysis (On et al. 2017), which is challenging and thus restricted to a small number of laboratories (Fox et al. 1989), is, however, likely to result in the publication of more robust and unambiguous new species descriptions within these genera. The main goal of taxonomy is to ensure investigators have the vocabulary to describe their group of organisms and be able to identify if they are related to organisms described by other investigators. Currently different species must be able to be differentiated by a phenotype and the term genomospecies can be used when there is a clear genetic difference between strains that can not yet be separated by a phenotype. In time, genomic taxonomy, which is currently in its infancy (Chun and Rainey 2014, Thompson et al. 2013), may provide robust species circumscription without the need for phenotypic markers of species delineation.

A range of bioinformatic software packages are also available for evaluating the pan-genome of a collection of genomes (Contreras-Moreira and Vinuesa 2013, Page et al. 2015,

Sahl et al. 2014). These packages can identify both the core genome of the collection as well as provide information of the accessory genome. The packages generally come with software that can be used to identify taxon-specific genes of a user-defined subset of genomes relative to a second subset. The taxon-specific genes identified can help identify pathways which may confer a phenotypic difference between the taxa represented in the analysis that can be exploited to define a new species. Significant effort is still required, however, to establish the function of many CDSs and characterise the resulting biochemical pathways to convert genotypic differences into a species-defining phenotype.These genes can also act as targets for taxon-specific molecular assays for taxa without existing assays. Alternatively these taxon-specific molecular assays can replace or complement existing assays based on conserved genes such as 16S rRNA or *hsp*60 which can suffer from cross-reaction to related species as the number of species within a genus increases.

6.3 Highly-multiplexed Detection of Microbial Taxa

Traditional methods for the identification of bacterial causes of human gastroenteritis relied on culture which requires a minimum of several days and up to a week or more for confirmed results. Technological advances have led to a growing number of multiplex molecular tests for syndromic gastrointestinal pathogen detection (Zhang, Morrison, & Tang, 2015). These kits typically detect less than 20 bacterial, protozoal and viral pathogens and the Epsilonproteobacterial taxa are limited to either generic Campylobacter spp. or specific *C. jejuni* detection. These kits provide a) reduced turnaround times, b) higher test specificity and sensitivity, c) improved co-infection detection, d) enhanced infection control, and e) more comprehensive testing relative to traditional methods (Zhang et al., 2015) but the inclusion of only the most common and well recognised pathogens limits the greater understanding of the causes of the large number of cases for which no aetiological agent is identified. Assays such as MLPA which offer greater multiplexing power, especially if coupled with next generation sequencing for detection and differentiation of the MLPA products (Benard-Slagter et al., 2017),

offer the opportunity to develop even more comprehensive, single test assays containing known as well as suspected pathogenic taxa.

These highly multiplexed MLPA assays are already used in human genetic disease diagnosis but they would also have application beyond the human clinical setting, including the testing of food and water samples, veterinary clinical testing, as well as evaluating environmental sustainability and ecological health. The targeted detection would provide better species and infraspecies discrimination than 16S rRNA sequencing and the sensitivity would be better than shotgun metagenomic analysis, at a reasonable cost. Thus highly multiplexed assays would provide a cost effective method for the detection of multiple known taxa in complex samples. Improvements would, however, need to be made in turn-around time and hands-on time to produce an assay that generates results within clinically useful timeframes (e.g. a single workday) without increasing the workload of technical staff.

Culture-independent diagnostic tests (CIDT), which include antigen- and DNA-based methods, are increasingly being used in clinical diagnosis of gastroenteritis (Marder et al. 2017). This has resulted in higher rates of positive stool samples but, since culture is often not performed as a reflex for many of these positive samples, fewer bacterial isolates are available for surveillance, outbreak investigation and antimicrobial susceptibility testing (Marder et al. 2017). Advances in sequencing technologies may soon provide sufficient DNA sequence information for these functions to be performed on data obtained directly from the stool samples. In the meantime, culture should be performed as a reflex test for CIDT-positive stool samples in order generate the isolates necessary for surveillance, outbreak investigation and antimicrobial susceptibility testing.

6.4 Future Directions

The genomic analysis described in Chapter 3 has provided valuable insights into the genomic taxonomy of the species *C. concisus* and the identification of GS-specific genes has been put to good use in the development of the MLPA described in Chapter 4. Next steps

include the identifying a phenotype that differentiates the two *C. concisus* genomospecies so that a new species (currently GS2) can be described. Closer interrogation of the pathways associated with the GS-specific genes identified here and by Chung et al. (2016) may provide clues to differences that confer a defining biochemical reaction that can be exploited to generate a phenotypic difference between the genomospecies.

The MLPA assay developed in Chapter 4 provides a rapid and cost effective method for detecting a broad range of Epsilonproteobacterial species in complex samples such as human stools. Future developments of the MLPA include replacing poor performing probes with alternatives. Additional whole genome sequences for some of the less well represented species will ensure the replacement probes are core to the taxon and absent, or highly divergent, from the pan-genome of other taxa. The size difference requirement of the MLPA assay, when using capillary electrophoresis for product detection, limits the number of probes that can be included to approximately 50 but this number can be increased if the MLPA products are detected by next generation sequencing (NGS) (Benard-Slagter et al. 2017, Kondrashova et al. 2015). Conversion of the Epsilonproteobacteria MLPA assay to this format would facilitate the inclusion of internal control probes, probes for the detection of additional pathogenic taxa, both within and beyond the Epsilonproteobacterial class, and the multiple probes per taxon. Future versions of the MLPA assay for clinical microbiology would offer direct detection of a broader range of known and suspected enteric pathogens including bacteria, viruses and parasites. MLPA coupled with NGS would provide a valuable tool for the detection of pathogenic microbes in complex samples but improvements in the turnaround time for both the MLPA and NGS components will be required to ensure the assay continues to provide results in a clinically useful timeframe. The portable real-time sequencers, such as the Oxford Nanopore MinION, may be able to provide a rapid solution to product detection but improvements in run cost, error rates and the types of errors generated need to be realised before this will be a viable option.

Chapter 5 provided some evidence that *C. concisus* GS2 is a risk factor for acute gastroenteritis in children. Additional epidemiological studies are required to confirm this finding

and to further the evidence towards inferring causation. Applying the MLPA assay to gastroenteritis case control studies involving other age groups would provide a better understanding of the demographics of *C. concisus* GS2 infection and provide a better understanding of the possible causative role of other Epsilonproteobacterial taxa, such as *C. ureolyticus*, in human gastroenteritis.

6.5 Conclusions

The additional eight *C. concisus* genomes generated in this project, and the genomic analysis undertaken on the collection of 22 genomes from this species, provide valuable information supporting the existence of two genomospecies. In addition, GS-specific genes were identified from this pan-genome with many of them also identified as GS-specific using an overlapping collection of genomes and different pan-genomic analysis software, which may provide clues to a phenotype that will allow the description of a new *Campylobacter* species.

An MLPA assay has been developed that provides a rapid and cost-effective tool for the simultaneous detection of 28 Epsilonproteobacterial taxa in a single assay, although further optimisation and the replacement of some probes is required. The utility of the MLPA assay was demonstrated by application to DNA extracts from faecal samples associated with a Belgian childhood gastroenteritis case control study. The MLPA results were consistent with *C. jejuni* as an important cause of gastroenteritis and also provided some evidence that *C. concisus* GS2 is a risk factor for childhood gastroenteritis.

Appendix I: Feature Frequency Profiling NeighborNets, Generated using a Range of *l*-mer Lengths, for 31 Genomes for *C. concisus* and Related Species

Ccon *C. concisus*, Ccur *C. curvus*, Cfet *C. fetus*, Chom *C. hominis*, Cjej *C. jejuni*, Cmuc *C. mucosalis*, Crec *C. rectus*, Csho *C. showae*, Cspu *C. sputorum*

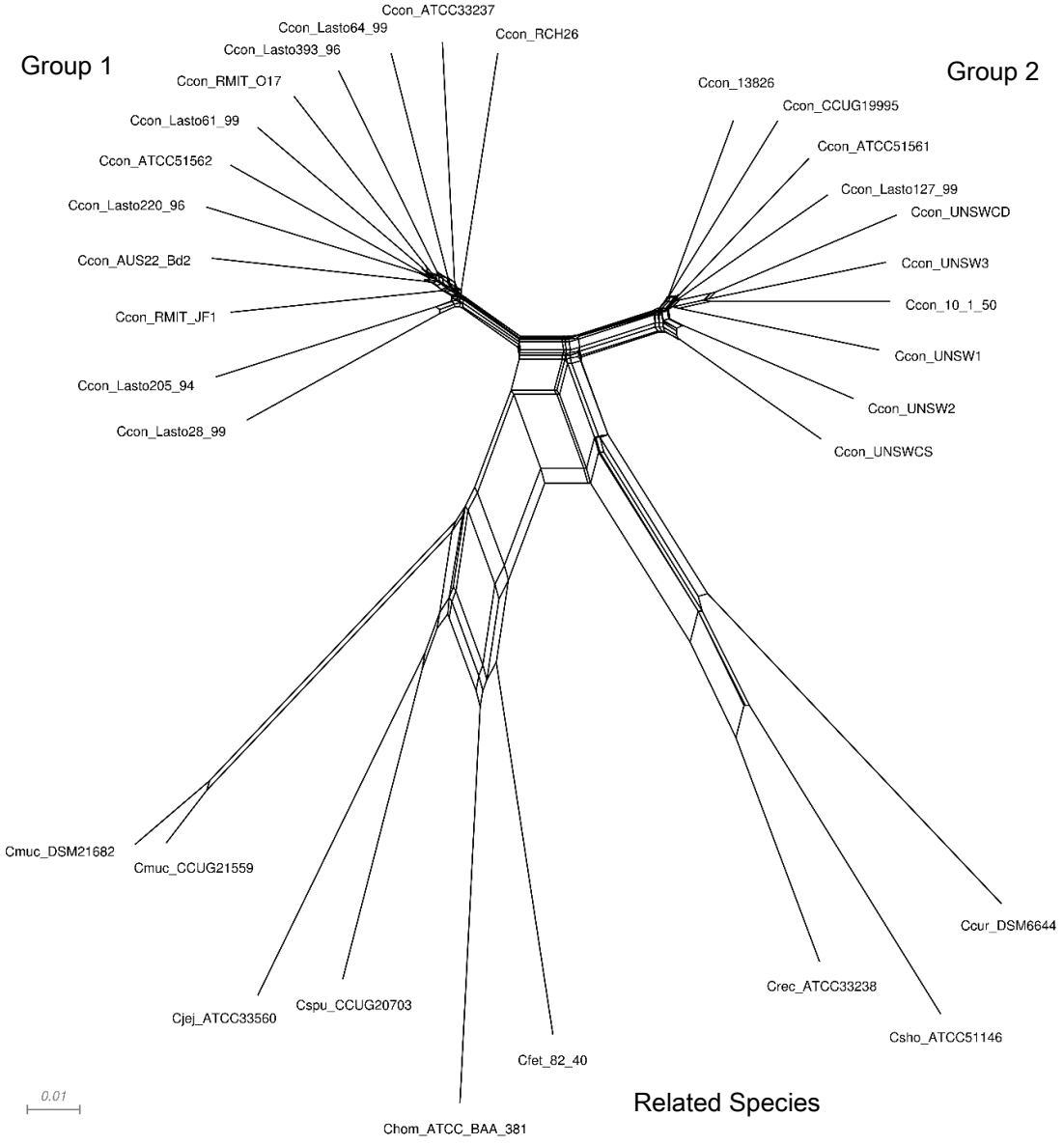


Figure 28: Feature Frequency Profiling NeighborNet (*l*-mer length of 11) for 31 Genomes representing *C. concisus* and Related Species

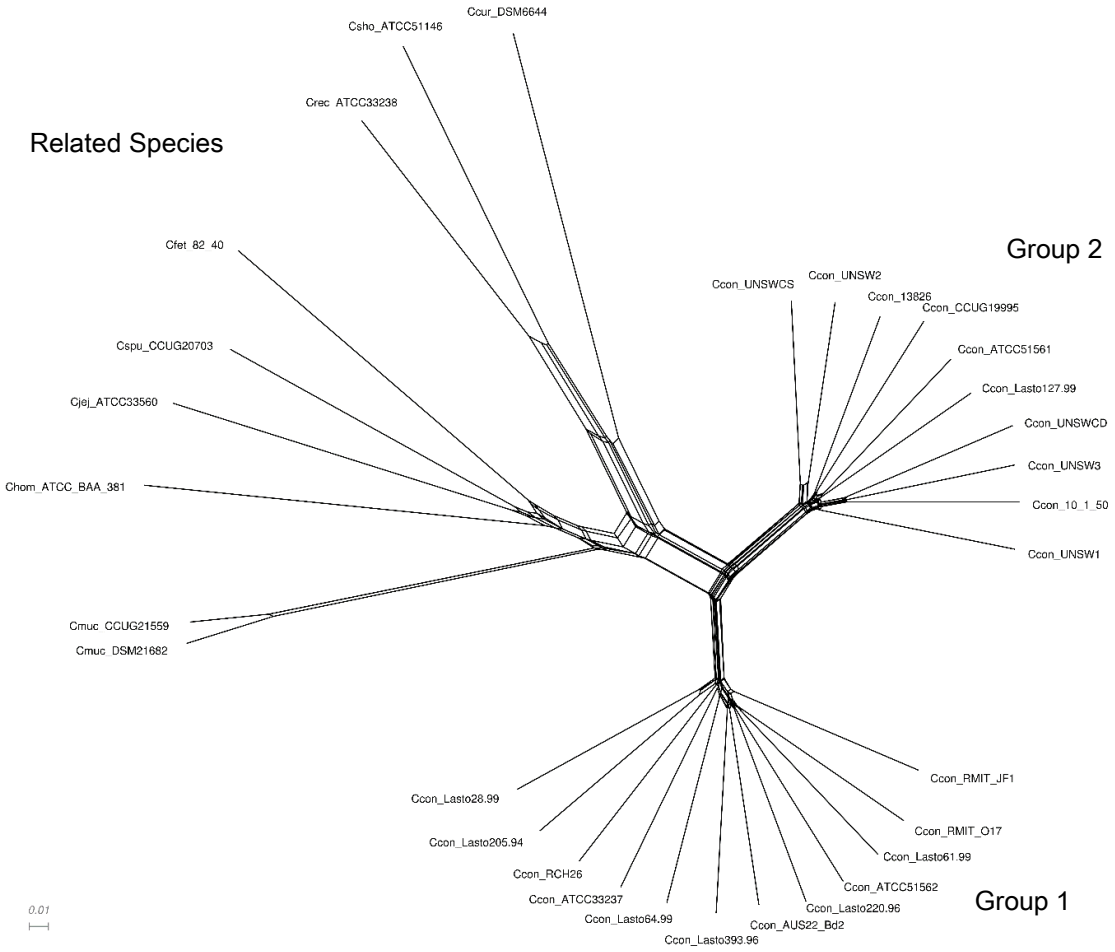


Figure 29: Feature Frequency Profiling NeighborNet (*l*-mer length of 12) for 31 Genomes representing *C. concisus* and Related Species

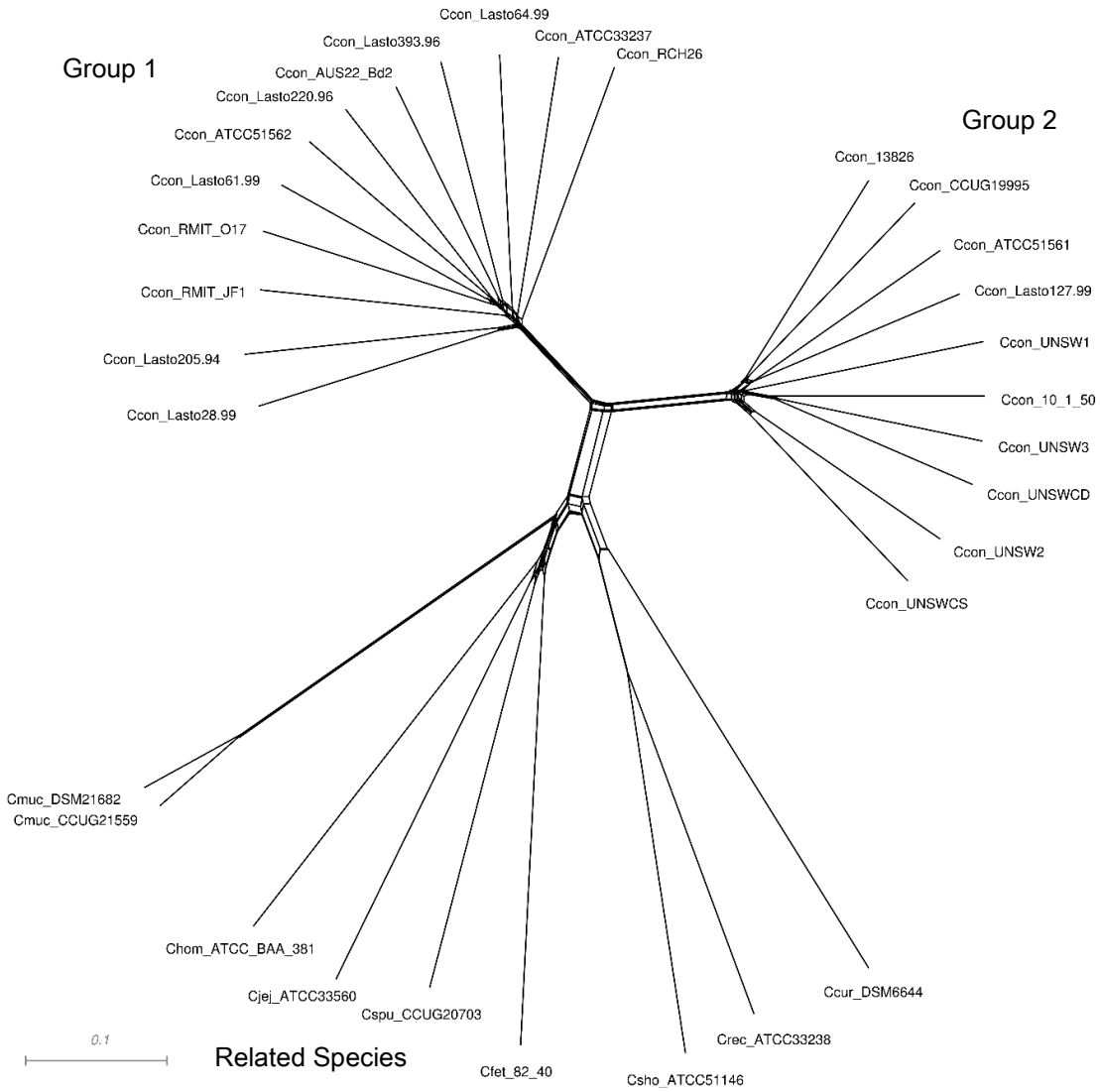


Figure 30: Feature Frequency Profiling NeighborNet (*l*-mer length of 13) for 31 Genomes representing *C. concisus* and Related Species

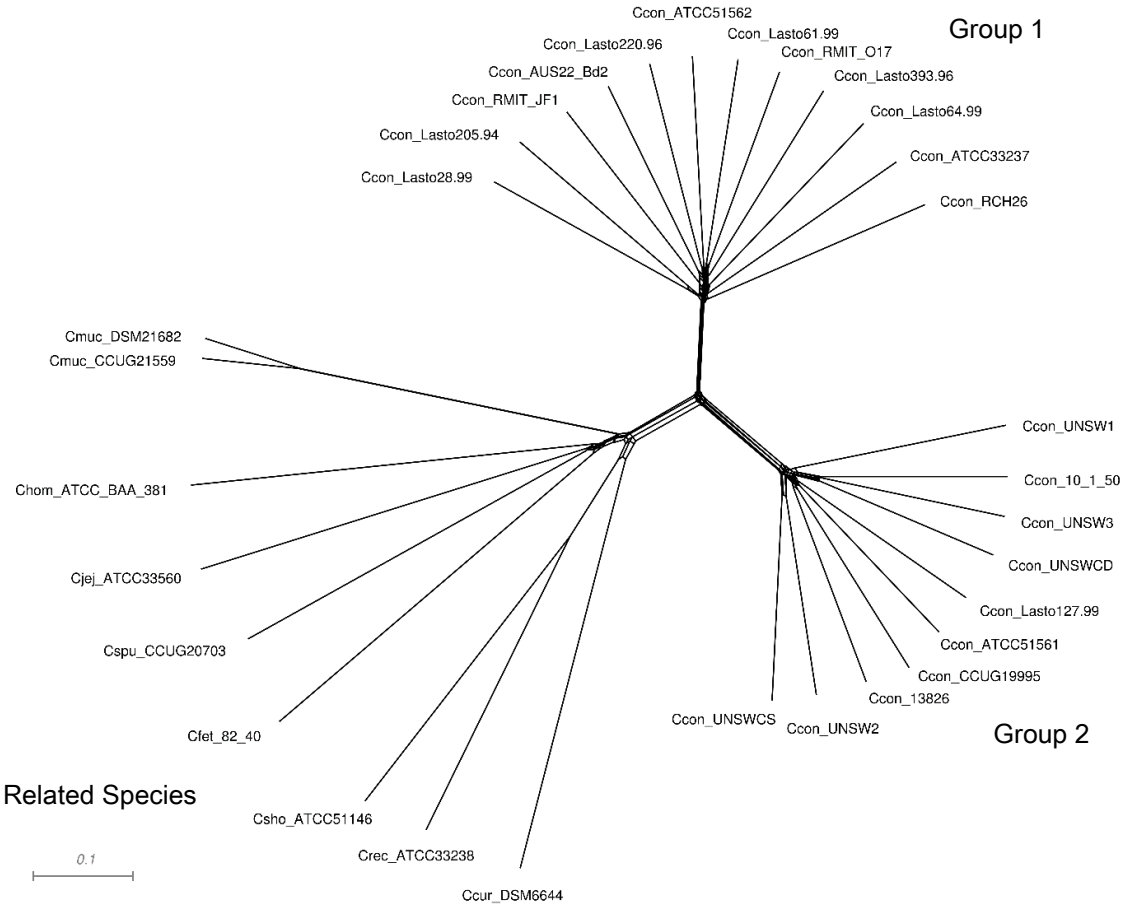


Figure 31: Feature Frequency Profiling NeighborNet (*l*-mer length of 14) for 31 Genomes representing *C. concisus* and Related Species

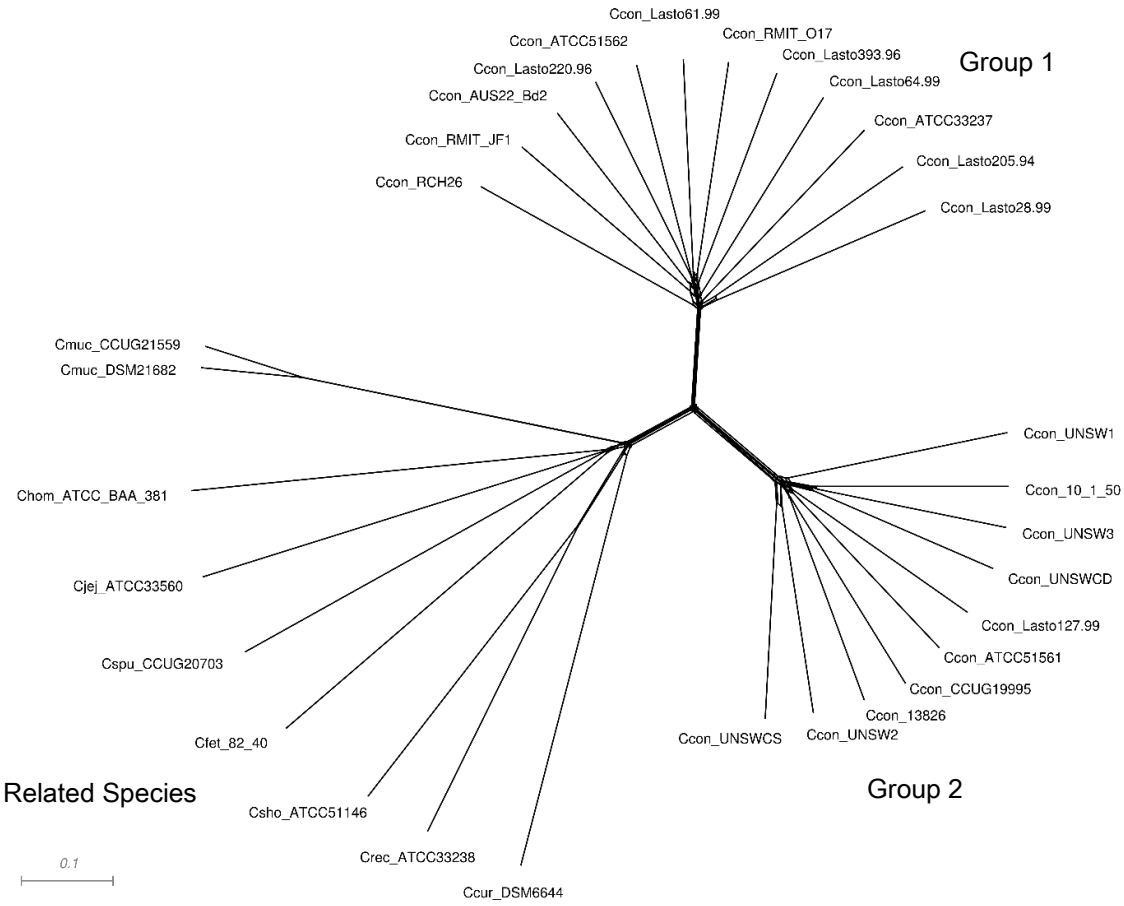


Figure 32: Feature Frequency Profiling NeighborNet (*l*-mer length of 15) for 31 Genomes representing *C. concisus* and Related Species

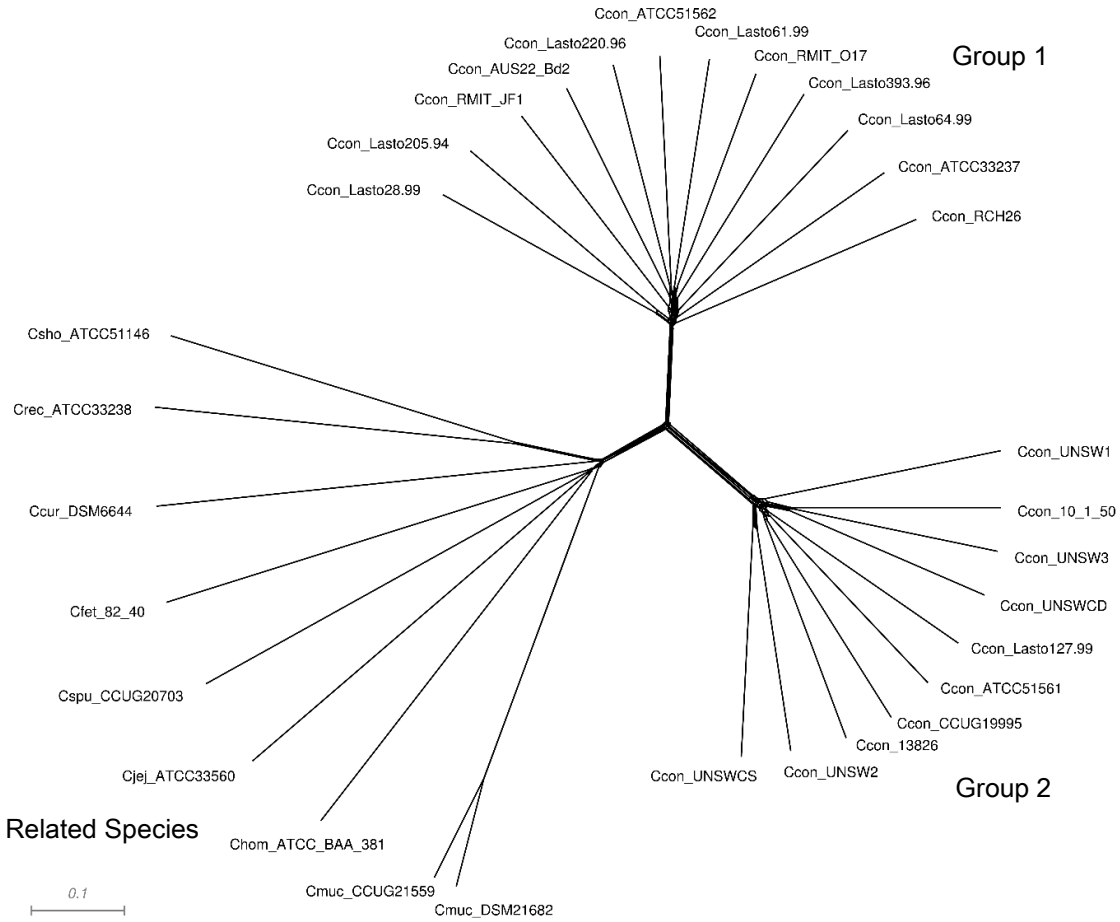


Figure 33: Feature Frequency Profiling NeighborNet (*l*-mer length of 16) for 31 Genomes representing *C. concisus* and Related Species

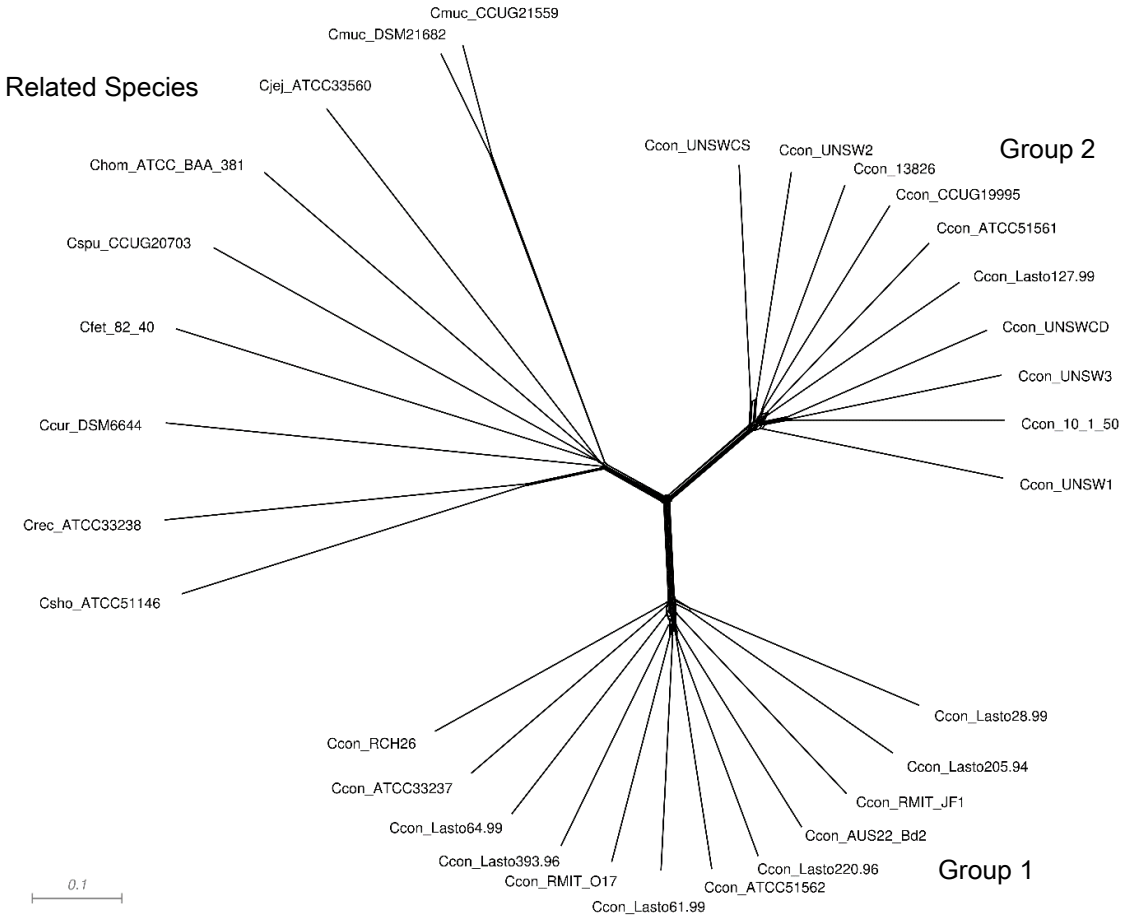


Figure 34: Feature Frequency Profiling NeighborNet (*l*-mer length of 17) for 31 Genomes representing *C. concisus* and Related Species

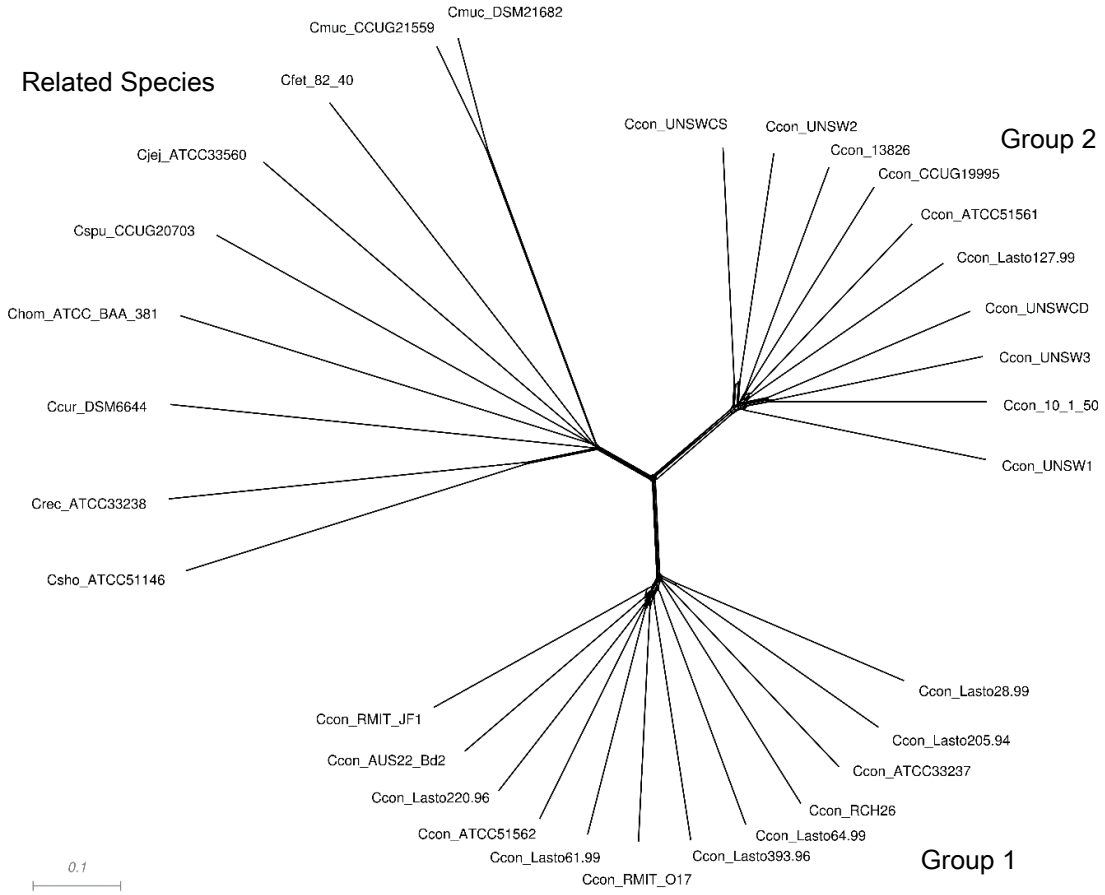


Figure 35: Feature Frequency Profiling NeighborNet (*l*-mer length of 18) for 31 Genomes representing *C. concisus* and Related Species

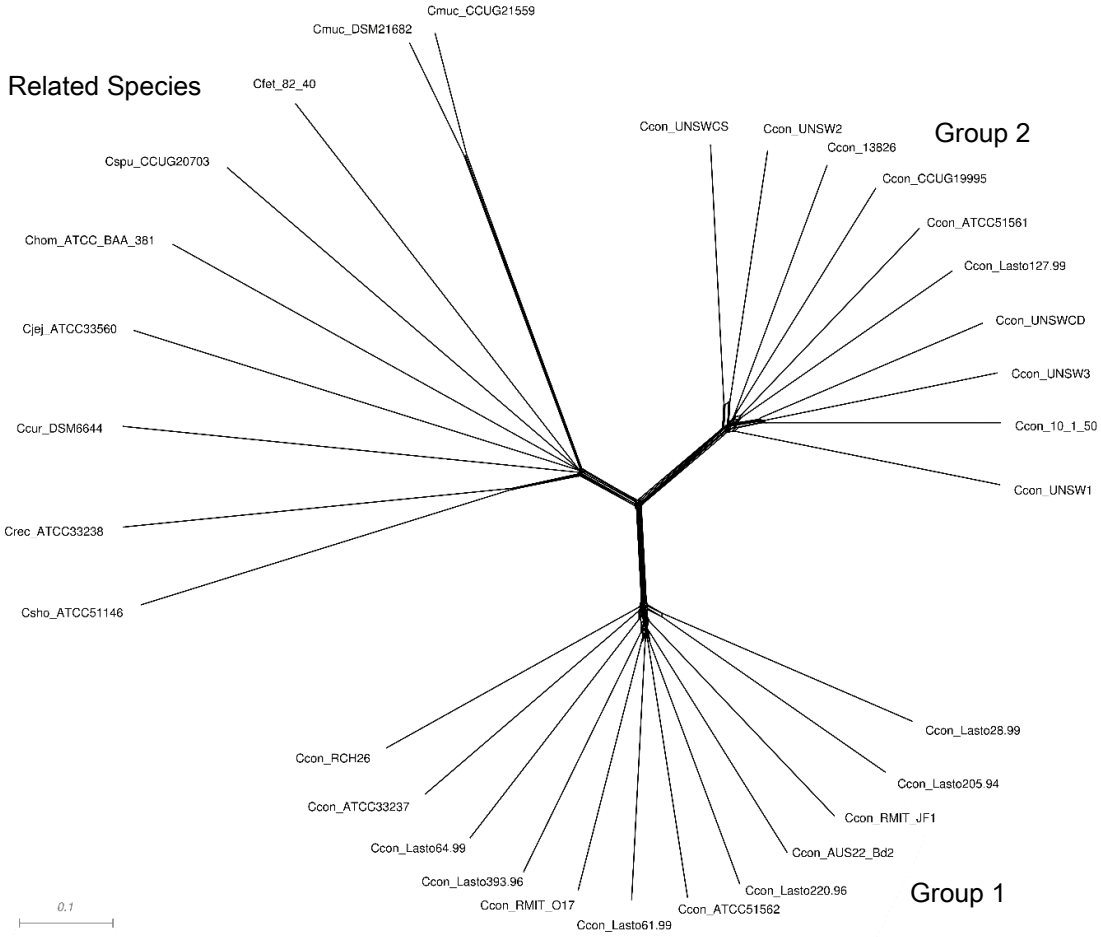


Figure 36: Feature Frequency Profiling NeighborNet (*l*-mer length of 19) for 31 Genomes representing *C. concisus* and Related Species

Appendix II: Ribosomal Multi-locus Sequence Typing (rMLST) Alleles for 22 *C. concisus* Genomes and nine Genomes Representing Related Species

Genes	Ccon_10_1_50	Ccon_13826	Ccon_ATCC33237	Ccon_ATCC51561	Ccon_ATCC51562	Ccon_AUS22_Bd2	Ccon_CCUG19995	Ccon_Lasto127_99	Ccon_Lasto205_94	Ccon_Lasto220_96	Ccon_Lasto28_99	Ccon_Lasto393_96	Ccon_Lasto61_99	Ccon_Lasto64_99	Ccon_RCH26	Ccon_RMIT_JF1	Ccon_RMIT_O17	Ccon_UNSW1	Ccon_UNSW2	Ccon_UNSW3	Ccon_UNSWCD	Ccon_UNSWCS	Ccur_DSM6644	Cfet_82_40	Chom_ATCC_BAA_381	Cjej_ATCC33560	Cmuc_CCUG21559	Cmuc_DSM21682	Crec_ATCC33238	Csho_ATCC51146	Cspu_CCUG20703
rplA	19	26	1	30	14	16	17	23	8	4	28	24	15	18	20	2	21	6	7	13	25	10	22	11	5	12	3	29	27	31	9
rplB	2	10	8	6	17	13	16	14	3	28	7	18	30	1	5	15	22	29	16	27	26	24	19	20	4	21	12	25	11	23	9
rplC	9	24	12	5	16	15	5	20	28	7	4	19	18	2	29	26	19	22	3	8	11	21	25	17	27	6	13	23	10	14	1
rplD	15	6	27	29	28	11	5	24	2	8	14	21	18	20	1	10	3	30	22	13	4	10	16	9	19	17	25	26	23	7	12
rplE	18	27	16	1	15	26	29	4	2	25	10	12	3	30	23	10	21	19	24	8	20	7	28	5	13	9	22	14	6	11	17
rplF	11	8	22	4	5	12	16	1	13	18	24	14	28	19	25	24	6	21	23	7	29	27	10	9	17	26	20	15	30	3	2
rplI	23	24	7	26	17	16	10	4	29	18	12	11	2	28	1	21	25	22	3	27	13	8	9	6	20	19	0	0	30	5	15
rplJ	10	21	13	19	6	5	25	8	27	1	4	2	24	28	26	12	15	9	8	11	30	7	20	14	17	18	3	22	29	23	0
rplK	27	5	3	11	20	25	19	6	10	4	1	21	17	25	12	2	14	23	24	29	7	22	8	13	26	9	15	15	18	16	28
rplL	27	8	16	28	9	7	4	2	11	29	24	25	3	12	15	21	23	19	13	31	10	5	26	14	1	6	22	20	17	30	18
rplM	13	8	21	13	20	15	10	14	1	24	4	18	9	15	3	11	1	13	23	25	10	13	6	2	16	12	19	17	5	22	7
rplN	11	3	14	7	14	24	18	10	25	26	22	23	17	1	11	22	14	6	19	3	27	8	13	4	5	2	20	16	9	15	21
rplO	20	2	15	14	23	9	4	6	22	28	29	8	16	24	26	29	30	27	5	18	21	7	13	3	10	19	12	25	1	11	17
rplP	19	30	18	15	11	4	1	10	24	16	21	27	8	13	29	28	12	9	2	7	26	20	3	6	17	22	14	5	23	31	25
rplQ	9	26	3	14	12	21	17	16	19	4	1	13	8	5	30	25	11	22	26	24	29	23	15	18	2	28	10	20	7	27	0

Genes	Ccon_10_1_50	Ccon_13826	Ccon_ATCC33237	Ccon_ATCC51561	Ccon_ATCC51562	Ccon_AUS22_Bd2	Ccon_CCUG19995	Ccon_Lasto127_99	Ccon_Lasto205_94	Ccon_Lasto220_96	Ccon_Lasto28_99	Ccon_Lasto393_96	Ccon_Lasto61_99	Ccon_Lasto64_99	Ccon_RCH26	Ccon_RMIT_JF1	Ccon_RMIT_O17	Ccon_UNSW1	Ccon_UNSW2	Ccon_UNSW3	Ccon_UNSWCD	Ccon_UNSWCS	Ccur_DSM6644	Cfet_82_40	Chom_ATCC_BAA_381	Cjej_ATCC33560	Cmuc_CCUG21559	Cmuc_DSM21682	Crec_ATCC33238	Csho_ATCC51146	Cspu_CCUG20703
rplR	1	5	12	13	20	10	19	11	4	16	24	28	18	21	14	15	27	31	25	8	17	3	26	30	9	23	6	7	22	29	2
rplS	28	24	25	17	1	27	5	26	16	4	11	9	8	9	6	14	21	7	23	19	22	30	13	15	29	3	10	18	2	20	12
rplT	22	24	13	28	2	1	18	25	9	27	8	11	12	7	16	13	9	3	4	15	17	26	21	10	20	6	23	19	5	5	14
rplU	9	17	22	10	23	3	8	25	5	16	19	5	14	6	15	18	7	21	26	12	9	26	27	2	11	4	24	1	13	28	20
rplV	21	3	23	5	15	4	16	12	11	9	20	19	18	25	2	18	6	21	7	5	5	16	8	24	14	22	0	0	1	17	10
rplW	14	6	17	6	9	6	6	6	6	9	13	6	3	10	11	6	16	6	4	4	6	6	2	1	7	12	0	0	18	5	15
rplX	13	17	6	6	16	22	5	6	3	3	20	5	7	2	10	20	21	14	6	19	14	6	12	1	11	8	0	0	18	4	15
rpmA	22	12	11	9	24	23	17	12	16	13	21	16	12	3	15	12	4	12	2	1	20	12	14	5	19	7	18	18	10	8	6
rpmB	14	4	3	20	16	16	22	19	23	16	3	11	5	2	3	23	21	13	18	19	17	19	1	6	12	8	15	15	7	9	10
rpmC	14	11	8	8	8	12	5	9	2	4	8	8	8	4	8	8	8	5	5	5	5	5	13	1	6	3	15	15	10	10	0
rpmD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rpmE	23	6	12	5	1	16	14	9	20	2	25	19	17	15	12	3	15	23	7	7	11	18	8	24	4	13	0	0	22	21	0
rpmF	5	5	1	5	14	2	5	5	13	12	13	18	15	6	1	2	14	5	11	5	5	5	16	4	3	17	10	10	7	9	8
rpmG	16	16	5	4	5	3	16	16	14	11	5	11	5	9	5	5	11	16	17	10	10	10	12	6	0	15	8	13	2	1	0
rpmH	1	19	8	7	8	16	19	11	16	16	4	16	9	10	4	12	5	1	7	7	19	1	2	13	14	15	3	18	17	20	0
rpmI	10	7	20	7	16	2	3	5	19	8	8	11	2	4	1	22	19	13	16	13	10	23	17	9	12	6	21	18	14	14	15
rpmJ	1	1	1	1	1	1	1	1	1	1	1	1	1	8	1	1	4	1	1	1	1	1	7	3	5	10	6	6	9	9	2
rpsA	3	31	6	30	29	23	21	1	13	7	10	4	18	25	11	19	12	9	27	16	17	8	5	28	2	15	20	22	26	24	14
rpsB	2	19	7	24	11	3	31	15	9	4	17	30	1	5	27	14	12	18	20	16	26	28	25	21	22	8	13	23	6	10	29
rpsC	15	26	8	25	30	10	31	5	24	2	13	21	4	11	12	29	19	1	18	20	22	17	14	28	7	27	9	3	23	16	6
rpsD	20	23	10	15	26	7	29	6	8	11	22	14	24	30	13	31	1	18	3	2	21	9	5	17	27	4	25	12	19	16	28

Genes	Ccon_10_1_50	Ccon_13826	Ccon_ATCC33237	Ccon_ATCC51561	Ccon_ATCC51562	Ccon_AUS22_Bd2	Ccon_CCUG19995	Ccon_Lasto127_99	Ccon_Lasto205_94	Ccon_Lasto220_96	Ccon_Lasto28_99	Ccon_Lasto393_96	Ccon_Lasto61_99	Ccon_Lasto64_99	Ccon_RCH26	Ccon_RMIT_JF1	Ccon_RMIT_O17	Ccon_UNSW1	Ccon_UNSW2	Ccon_UNSW3	Ccon_UNSWCD	Ccon_UNSWCS	Ccur_DSM6644	Cfet_82_40	Chom_ATCC_BAA_381	Cjej_ATCC33560	Cmuc_CCUG21559	Cmuc_DSM21682	Crec_ATCC33238	Csho_ATCC51146	Cspu_CCUG20703
rpsE	18	14	5	29	25	4	30	19	27	23	8	3	24	15	17	22	2	1	11	20	10	21	9	7	28	12	13	6	31	16	26
rpsF	4	20	11	25	24	21	28	26	17	1	16	13	14	3	6	15	7	23	19	30	29	2	8	22	10	27	0	0	5	9	18
rpsG	17	5	29	10	18	26	14	19	27	13	11	4	15	24	20	1	30	25	23	3	8	14	21	12	9	2	16	22	7	6	28
rpsH	7	15	16	29	17	19	20	10	26	14	2	21	23	24	5	2	28	4	13	6	8	27	30	11	3	25	1	12	22	18	9
rpsI	13	23	19	11	5	26	27	25	12	20	7	21	6	3	16	26	4	25	14	17	25	22	15	24	9	2	10	10	18	8	1
rpsJ	6	6	5	6	4	6	6	6	6	6	6	6	6	6	6	10	6	6	7	6	6	6	1	3	13	9	12	12	8	11	2
rpsK	17	3	27	15	11	23	1	18	4	8	9	20	6	16	21	14	13	18	29	5	30	19	12	10	28	7	2	24	25	26	22
rpsL	7	15	21	23	22	3	10	15	14	24	5	3	16	12	13	1	18	15	15	15	15	15	8	11	9	4	19	17	20	2	6
rpsM	9	1	22	3	25	7	29	24	4	18	22	21	14	19	5	28	19	12	26	2	27	23	8	17	20	11	16	13	15	10	6
rpsN	9	9	6	9	8	8	9	9	8	13	8	7	10	2	15	8	2	9	9	9	9	9	3	12	17	14	4	5	1	16	11
rpsO	3	3	14	3	11	7	3	3	21	17	6	3	17	3	16	5	4	12	3	3	18	2	1	19	9	8	13	13	15	20	10
rpsP	26	5	15	16	25	1	23	29	19	1	2	1	20	11	12	4	21	27	10	22	9	24	28	8	3	14	6	17	13	18	0
rpsQ	7	21	3	1	14	2	25	12	6	22	11	17	5	3	7	11	3	20	10	16	1	19	24	18	23	15	9	9	13	26	8
rpsR	13	16	20	25	6	22	8	4	14	1	7	14	27	22	29	18	6	5	28	2	3	24	21	26	9	10	11	15	17	12	23
rpsS	5	5	10	5	5	5	5	5	10	8	11	5	5	5	5	11	5	5	5	5	5	5	6	2	7	1	9	4	13	3	12
rpsT	5	18	28	29	1	19	22	10	11	23	14	13	3	27	11	6	24	16	2	21	9	9	15	20	8	4	25	17	7	26	0
rpsU	22	13	16	1	13	21	11	17	22	9	22	9	22	13	3	4	19	2	15	14	20	7	12	23	24	10	18	18	8	5	6

Ccon *C. concisus*, Ccur *C. curvus*, Cfet *C. fetus*, Chom *C. hominis*, Cjej *C. jejuni*, Cmuc *C. mucosalis*, Crec *C. rectus*, Csho *C. showae*, Cspu *C. sputorum*

Appendix III: Generation of Control DNA

Materials and Methods

A selection of 97 strains representing 43 species within the *Arcobacter, Campylobacter*, and *Helicobacter* genera were grown on Columbia Horse Blood agar (CBA, Fort Richard, Auckland, New Zealand) or Tryptic Soy Agar (TSA, Fort Richard, Auckland, New Zealand) as described in Table 19. A selection of 15 strains, representing commonly reported Gram positive and Gram negative human bacterial pathogens, were also included in the study so that the specificity of the MLPA probes could be evaluated beyond the Epsilonproteobacterial class. The growth conditions for these strains are also summarised in Table 19. Four of the *A. butzleri* strains (A758, BRUG0093, BRUG0198 and BRUG0318) and one of the *C. upsaliensis* strains (F221) were kindly provided by Professor Olivier Vandenberg (St. Pierre University Hospital, Belgium). Cultures of two well-characterised *H. pylori* strains were prepared by Dr Jacqueline Keenan (University of Otago). These *H. pylori* strains had been grown at 37°C in 6-12% O_2, 5-8% CO_2 as generated by a MicroAero gas generation pack (Mitsubishi Gas Chemical [MGC], Tokyo, Japan). The remaining strains requiring microaerobic conditions were grown in a MAC-VA 500 cabinet (Don Whitely Scientific, Shipley, United Kingdom) with an atmosphere of 10% CO_2, 7% H_2, 3% O_2 and 80% N_2 and the strains requiring anaerobic conditions were grown in an atmosphere containing 0% O_2 and 16% CO_2, the atmosphere generated by a Anaero gas generation pack (Mitsubishi Gas Chemical). For strains unable to be recovered from frozen storage, DNA was extracted directly from the frozen culture.

Slightly turbid suspensions were prepared in 1 mL of Phosphate Buffered Saline (BR0014G, Oxoid, Basingstoke, England) and the DNA was extracted using the Blood and Tissue Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. The elution buffer AE has the composition 10 mM Tris-Cl, 0.5 mM EDTA, pH 9.0. RNaseI treatment was included for all extracts except the two *H. pylori* strains and the 15 non-Epsilonproteobacteria

human enteric pathogens. Ethanol precipitation was performed on some extracts where the concentration of DNA extracted was below 20 ng/µL. This was achieved by mixing the DNA with 0.1 volumes of 3M sodium acetate pH 5.2 and then adding twice the volume of absolute ethanol. The samples were held at -80°C overnight and then centrifuged at 16,000 x *g* for 20 min. The supernatant was discarded, the pellet air-dried and resuspended in 20 µL low EDTA-TE buffer (10 mM Tris [pH 8.0], 0.1 mM EDTA, 12090-015, Invitrogen, Carlsbad, USA).

An additional 45 DNA samples representing 28 species, as described in Table 20, were provided by Dr William Miller (United States Department of Agriculture, USA), Dr Francis Megraud (University of Bordeaux, France), Dr Maria Figueras (Universitat Rovira i Virgili, Spain), Professor Olivier Vandenberg (St. Pierre University Hospital, Belgium), Dr Anne Midwinter and Dr Krunoslav Bojanic (Massey University). Extracts received dried were resuspended in sufficient dH_2O to saturate the paper they were dried on and provide 100 µL of DNA (at least 500 µL). Ethanol precipitation was undertaken on DNA where the concentration was below 20 ng/µL. The method was the same as above but the air-dried pellet was resuspended in 60 µL of dH_2O.

The DNA concentration and quality was estimated using a Nanodrop 1000 (Thermo Scientific, Waltham, MA) and a more accurate concentration established using a Qubit (Life Technologies, Carlsbad, USA). Stock solutions at 20 ng/µL from ESR were generally prepared in high EDTA-TE buffer (10 mM Tris [pH 8.0], 1 mM EDTA, in house). In contrast, for collaborators DNA of sufficiently high concentration, stock solutions at 20 ng/µL were prepared in low EDTA-TE buffer (12090-015, Invitrogen; 10 mM Tris [pH 8.0], 0.1 mM EDTA). Additional solutions, at lower concentrations, of DNA extracted from some strains, as well as mixtures of DNA, were also prepared in low EDTA-TE buffer.

16S rRNA gene sequences were generated for at least one DNA extract from at least one strain representing each target taxa, where possible, to check the identity of the DNA. Briefly 2 µL of DNA was amplified in a 25 µL reaction containing 1 X AmpliTaq Gold® 360 Master Mix (ThermoFisher, Waltham, MA) and 200 µM of each of 16F27 and 16F1541 Table 18. The

thermal profile consisted of an initial denaturation of 5 min at 95°C, 40 cycles of 15 sec at 94°C, 30 sec at 60°C and 30 sec at 72°C, and a final extension of 7 min at 72°C. The products were sequenced on an ABI genetic analyser 3130XL using the primers listed in Table 18. The sequences were *de novo* assembled in Geneious (R9.1.7) to produce almost full length gene sequences, in most cases. These sequences were used as the query for searches of the local and/or NCBI nr databases within Geneious (R9.1.7) using BLASTn, a word size of 11 and a maximum of 100 hits.

Table 18: Primers Used for 16S rRNA Sequencing of Control DNA

Primer Name	Sequence	Used For
16F27	5'-AGA GTT TGA TCM TGG CTC AG-3'	PCR and sequencing
16F49	5'-TAA YAC ATG CAA GTC GA-3'	Sequencing
16F357	5'-ACT CCT ACG GGA GGC AGC AG-3'	Sequencing
16F530	5'-GTG CCA GCM GCC GCG G-3'	Sequencing
16F945	5'-GGG CCC GCA CAA GCG GTG G-3'	Sequencing
16R518	5'-CGT ATT ACC GCG GCT GCT GG-3'	Sequencing
16R1087	5'-CTC GTT GCG GGA CTT AAC CC-3'	Sequencing
16R1389	5'-ACG GGC GGT GTG TAC AAG-3'	Sequencing
16R1492	5'-TAC GGY TAC CTT GTT ACG ACT T-3'	Sequencing
16R1541	5'-AAG GAG GTG ATC CAG CCG CA-3'	PCR and sequencing

DNA extracts with 16S rRNA-based identifications that different from what was expected were tested using the Epsilonproteobacteria MLPA assay as described in Chapter 4 and the results were excluded from the evaluation of the assay if the MLPA results confirmed the 16S rRNA results.

Results

A total of 97 DNA extracts, representing 41 species, were prepared to evaluate the specificity of the Epsilonproteobacteria MLPA (Table 19). An additional 44 DNA extracts, representing 28 species, were provided by collaborators as summarised in Table 20.

.

Table 19: Summary of Control DNA Extracts Prepared to Evaluate the Specificity of the Epsilonproteobacteria Multiplex Ligation-Dependent Probe Amplification (MLPA) Assay

The Best Hits from both the local EpsiloFsa and NCBI nr databases were included for Epsilonproteobacteria taxa and only the NCBI nr database for other taxa. The Best Hit was the Taxon for which the highest Grade was recorded. Grade is a percentage calculated by Geneious from the query coverage, e-value and identity and aides in the identification of the longest, highest identity hits.

Taxa	ATCC[a]	CCUG[b]	NCTC[c]	NZRM[d]	Other[e]	Growth conditions	Length (nt)	EpsiloFsa Best Hit	nr Best Hit
A. butzleri	49616^T	**30485^T**	12481^T	4017^T	RM 4018^T	Aerobic at 25°C on CBA[f]	1464	*A. butzleri* Grade 100%	*A. butzleri* Grade 100%
A. butzleri					**A758**	Aerobic at 25°C on CBA	1468	*A. butzleri* Grade 100%	*A. butzleri* Grade 100%
A. butzleri					**BRUG0093**	Aerobic at 25°C on CBA	1470	*A. butzleri* Grade 100%	*A. butzleri* Grade 100%
A. butzleri					**BRUG0198**	Aerobic at 25°C on CBA	1467	*A. butzleri* Grade 100%	*A. butzleri* Grade 100%
A. butzleri					**BRUG0318**	Aerobic at 25°C on CBA	1467	*A. butzleri* Grade 100%	*A. butzleri* Grade 100%
A. cryaerophilus	43158^T	**17801^T**	11885^T	4018^T	RM 1582^T	Aerobic at 25°C on CBA	1487	*A. cryaerophilus* Grade 100%	*A. butzleri* Grade 98.7% (40 SNP, 1487 nt of 1509 nt partial sequence), *A. cryaerophilus* Grade 98.5% (1 SNP, 3 N in hit, 1445 nt of 1462 nt partial sequence)
A. skirrowii	51132^T	**10374^T**			RM 3222^T	Aerobic at 25°C on CBA	1480	*A. skirrowii* Grade 100%	*A. skirrowii* Grade 98.7% (1 SNP, 1 Y in hit, 1444 nt of 1462 nt partial sequence)
Bacillus cereus	10702		8035	**5**		Aerobic at 37°C on CBA	1442		*B. anthracis/B. toyonensis/B. thuringiensis/B. cereus* Grade 100%

Taxa	ATCC[a]	CCUG[b]	NCTC[c]	NZRM[d]	Other[e]	Growth conditions	Length (nt)	EpsiloFsa Best Hit	nr Best Hit
C. canadensis		**54429[T]**				MAC[g] at 37°C on CBA	916 & 461	*C. canadensis* Grade 99.9% (1 SNP) & *C. canadensis* Grade 100%	*C. canadensis* Grade 98.2% (1 insertion, 885 nt of 1361 nt partial sequence) & *C. canadensis* Grade 100%
C. coli	33559[T]	**11283[T]**	11366[T]	2607[T]	RM 1875[T]	MAC at 37°C on CBA	1417	*C. coli* Grade 100%	*C. coli* Grade 100%
C. coli					CMB06690	MAC at 37°C on CBA			
C. coli					CMB091229	MAC at 37°C on CBA			
C. coli					CPH0311750	MAC at 37°C on CBA			
C. coli					RM 2228	MAC at 37°C on CBA			
C. concisus GS1	33237[T]	**13144[T]**	11485[T]		RM 7084[T]	MAC at 37°C on CBA	1481	*C. concisus* Grade 99.9%	*C. concisus* Grade 100%
C. concisus GS1					**Lasto115.99**	MAC at 37°C on CBA			
C. concisus GS1					**Lasto220.96**	MAC at 37°C on CBA			
C. concisus GS1					**Lasto24.99**	MAC at 37°C on CBA			
C. concisus GS1					**Lasto28.99**	MAC at 37°C on CBA			
C. concisus GS1					**Lasto389.96**	MAC at 37°C on CBA			
C. concisus GS1					**Lasto61.99**	MAC at 37°C on CBA			

Taxa	ATCC[a]	CCUG[b]	NCTC[c]	NZRM[d]	Other[e]	Growth conditions	Length (nt)	EpsiloFsa Best Hit	nr Best Hit
C. concisus GS1					**Lasto64.99**	MAC at 37°C on CBA			
C. concisus GS2		**19995**				MAC at 37°C on CBA	1479	*C. concisus* Grade 100%	*C. concisus* Grade 99.9% (4 SNP)
C. concisus GS2					**Lasto104.93**	MAC at 37°C on CBA			
C. concisus GS2					**Lasto113.99**	MAC at 37°C on CBA			
C. concisus GS2					**Lasto131.99**	MAC at 37°C on CBA			
C. concisus GS2					**Lasto135.99**	MAC at 37°C on CBA			
C. concisus GS2					**Lasto140.99**	MAC at 37°C on CBA			
C. concisus GS2					**Lasto275.95**	MAC at 37°C on CBA			
C. concisus GS2					**Lasto316.98**	MAC at 37°C on CBA			
C. concisus GS2					**Lasto377.96**	MAC at 37°C on CBA			
C. cuniculorum		**56289^T**			DSM 23162^T, RM 8641^T	MAC at 37°C on CBA	1483	*C. cuniculorum* Grade 100%	*C. jejuni* Grade 98.1% (55 SNP, 1483 nt of complete gene), *C. cuniculorum* Grade 97.6%, (0 SNP, 1411 nt of 1414 nt partial sequence)
C. curvus	35224^T	**13146^T**	11649^T		DSM 6644^T	Anaerobic[h] at 37°C on CBA	1566	*C. curvus* Grade 100%	*C. curvus* Grade 99.5% (3 SNP, 1552 nt of 1555 nt partial sequence)

Taxa	ATCC[a]	CCUG[b]	NCTC[c]	NZRM[d]	Other[e]	Growth conditions	Length (nt)	EpsiloFsa Best Hit	nr Best Hit
C. fetus subsp. *fetus*	27374^T	**6823^T**	10842^T	2398^T		MAC at 37°C on CBA	1449	*C. fetus* (subspp. *fetus* and *venerealis*) Grade 100%	*C. fetus* (subspp. *fetus* and *venerealis*) Grade 100%
C. fetus subsp. *venerealis*	19438^T	**538^T**	10354^T	2399^T		MAC at 37°C on CBA	1437	*C. fetus* (subspp. *fetus* and *venerealis*) Grade 100%	*C. fetus* (subspp. *fetus* and *venerealis*) Grade 100%
C. gracilis	33236^T	**27720^T**	12738^T		RM 3268^T	Anaerobic at 37°C on CBA	1426	*C. gracilis* Grade 100%	*C. gracilis* Grade 100%
C. helveticus		**30563**				MAC at 37°C on CBA			
C. helveticus	51209^T	**30682^T**	12470^T		RM 3228^T	MAC at 37°C on CBA	1464	*C. helveticus* Grade 100%	*C. helveticus* Grade 99.1% (1 deletion, 1 SNP, 1439 nt of 1439 nt sequence)
C. helveticus		**30683**	12471			MAC at 37°C on CBA			
C. hominis	BAA-381^T	45161^T	13146^T		**CH001^T**	Anaerobic at 37°C on CBA	1463	*C. hominis* Grade 100%	*C. hominis* Grade 100%
C. hyointestinalis subsp. *hyointestinalis*	35217^T	**14169^T**	11608^T	3676^T	DSM 19053^T	MAC at 37°C on CBA	1445	*C. hyointestinalis* subsp. *hyointestinalis* Grade 100%	*C. hyointestinalis* subsp. *hyointestinalis* Grade 100%
C. hyointestinalis subsp. *lawsonii*		**27631**			RM 4096	MAC at 37°C on CBA	1426	*C. hyointestinalis* subsp. *lawsonii* Grade 100%	*C. hyointestinalis* subsp. *lawsonii* Grade 100%
C. insulaenigrae		48653^T	12927^T		**LMG 22716^T,** RM 5435^T	MAC at 37°C on CBA	981 & 460	*C. insulaenigrae* Grade 99.9% (1 insertion) & *C. insulaenigrae* Grade 100%	*C. insulaenigrae* Grade 99.9% (1 insertion) & *C. insulaenigrae* Grade 100%

Taxa	ATCC[a]	CCUG[b]	NCTC[c]	NZRM[d]	Other[e]	Growth conditions	Length (nt)	EpsiloFsa Best Hit	nr Best Hit
C. jejuni subsp. *doylei*	49349^T	**24567^T**	11951^T	3516^T		MAC at 37°C on CBA	880 & 458	*C. jejuni* subsp. *doylei* Grade 100% & *C. jejuni* (including *C. jejuni* subsp. *doylei*)| *C. coli* Grade 100%	*C. jejuni* subsp. *doylei* Grade 100% & *C. jejuni* (including *C. jejuni* subsp. *doylei*)| *C. coli* Grade 100%
C. jejuni subsp. *jejuni*	33560^T	**11284^T**	11351^T	2397^T		MAC at 37°C on CBA	1419	*C. jejuni* subsp. *jejuni* Grade 100%	*C. jejuni* subsp. *jejuni* Grade 100%
C. jejuni subsp. *jejuni*			**11168**			MAC at 37°C on CBA			
C. jejuni subsp. *jejuni*					**RM 1221**	MAC at 37°C on CBA			
C. jejuni subsp. *jejuni*					**RM 1864** = 81-176	MAC at 37°C on CBA			
C. jejuni subsp. *jejuni*					**P110B**	MAC at 37°C on CBA			
C. lanienae		44467^T	**13004^T**		RM 3663^T	MAC at 37°C on CBA	1446	*C. lanienae* Grade 100%	*C. lanienae* Grade 99.7%
C. lari subsp. *lari*	35221^T	**23947^T**	11352^T	2622^T		MAC at 37°C on CBA	1464	*C. lari* Grade 100%	*C. lari* Grade 100%
C. mucosalis	43264^T	**6822^T**	11000^T		DSM 21682^T	MAC at 37°C on CBA	1463	*C. mucosalis* Grade 100%	*C. mucosalis* Grade 99.1% (1 insertion, 1 deletion, 17 SNP, 1457 nt of 1468 nt partial sequence)
C. peloridis		55787^T			**R-13342^T**, LMG 23910^T, RM 14092^T	MAC at 37°C on CBA	733 & 612	Both *C. peloridis* Grade 100%	Both *C. peloridis* Grade 100%
C. rectus	33238^T	**20446^T**	11489^T		RM 6916^T	Anaerobic at 37°C on CBA	1487	*C. ureolyticus* Grade 100%	*C. ureolyticus* Grade 99.9% (4 SNP, 1487 nt of 1501 partial sequence)

Taxa	ATCC[a]	CCUG[b]	NCTC[c]	NZRM[d]	Other[e]	Growth conditions	Length (nt)	EpsiloFsa Best Hit	nr Best Hit
C. rectus	**33238[T]**	20446[T]	11489[T]		RM 6916[T]	Anaerobic at 37°C on CBA			
C. showae	51146[T]	**30254[T]**			RM 3277[T]	Anaerobic at 37°C on CBA	1469	*C. showae* Grade 100%	*C. rectus* Grade 99.0% (29 SNP, 1469 nt of 1485 nt partial sequence), *C. showae* Grade 99.7% (2 SNP, 1440 nt of 1440 nt partial sequence)
C. sputorum biovar *sputorum*	33562	**11289**	11367			MAC at 37°C on CBA	232	*C. sputorum* (biovar *sputorum*, *fecalis* & *paraureolyticus*) Grade 99.4% (229 nt of complete gene, SNP last 3 nt)	*C. sputorum* Grade 99.4% (229 nt of 1695 nt partial sequence)
C. sputorum biovar *sputorum*	35980[T]	**9728[T]**	11528[T]		CIP 103749[T]	MAC at 37°C on CBA		Sequence too poor to generate sequence for BLAST searching	
C. subantarcticus		**38513[T]**			LMG 24377[T], RM 8523[T]	MAC at 37°C on CBA	1448	*C. subantarcticus* Grade 100%	*C. subantarcticus* Grade 100%
C. upsaliensis	43954[T]	14913[T]	11541[T]	**3675[T]**	DSM 5365[T]	MAC at 37°C on CBA	1428	*C. upsaliensis* Grade 99.9% (1 N in query sequence, SNP first 2 nt)	*C. upsaliensis* Grade 99.3% (6 SNP including first 2 nt of query, 1 N in query sequence, 1426 nt of 1435 nt partial sequence)
C. upsaliensis					**F221**	MAC at 37°C on CBA	960 & 398	*C. upsaliensis* Grade 99.4% (11 SNP) & *C. upsaliensis* Grade 99.6% (3 SNP)	*C. upsaliensis* Grade 99.9% (1 SNP, 960 nt of 1417 partial sequence) & *C. upsaliensis* Grade 99.6% (3 SNP, 398 nt of 1417 nt partial sequence)
UPTC		**20707**				MAC at 37°C on CBA	1496	UPTC Grade 100%	UPTC Grade 100%

Taxa	ATCC[a]	CCUG[b]	NCTC[c]	NZRM[d]	Other[e]	Growth conditions	Length (nt)	EpsiloFsa Best Hit	nr Best Hit
C. ureolyticus	33387^T	**7319^T**	10941^T		DSM 20703^T	Anaerobic at 37°C on CBA	1476	*C. ureolyticus* Grade 100%	*C. ureolyticus* Grade 99.9% (4 SNP, 1476 nt of 1501 nt partial sequence)
Clostridium difficile	9689^T	4938^T	11209^T	**2390^T**		Anaerobic at 37°C on CBA	1389		*C. difficile* Grade 100%
Clostridium perfringens	13124^T	1795^T	8237^T	**20^T**		Anaerobic at 37°C on CBA	1408		*C. perfringenes* Grade 100%
Escherichia coli	**11775^T**	24^T	9001^T	3309^T		Aerobic at 37°C on TSA	1492		*E. coli* Grade 100%
Escherichia coli O157	700728		12900	**3614**		Aerobic at 37°C on CBA	1411		*E. coli* (including O157) Grade 100%
H. acinonychis	**51101^T**	29263^T	12686^T			MAC at 37°C on CBA			
H. aurati	BAA-1^T	**47791^T**				MAC at 37°C on CBA			
H. canadensis	700968^T	**47163^T**	13241^T		MIT 98-5491^T	MAC at 37°C on CBA	1460	*H. canadensis* Grade 100%	*H. canadensis* Grade 99.9% (0 SNP, 1457 nt of 1457 nt partial sequence)
H. canis	51401^T	**32756^T**	12739^T			MAC at 37°C on CBA	1463	*H. canis* Grade 99.9% (2 SNP including last nt, 1462 nt of complete gene)	*H. canis* Grade 99.9% (0 SNP, 1460 nt of 1473 nt partial sequence)
H. cinaedi	BAA-847^T	**18818^T**	12423^T			MAC at 37°C on CBA	1449	*H. cinaedi* Grade 100%	*H. cinaedi* Grade 100%
H. fennelliae	35684^T	**18820^T**	11612^T			MAC at 37°C on CBA	1786	*H. fennelliae* Grade 100%	*H. fennelliae* Grade 99.9% (Pairwise Identity & Identical Sites 100%, 1784 nt)
H. hepaticus	51448^T	**33637^T**				MAC at 37°C on CBA	1443	*H. hepaticus* Grade 100%	*H. hepaticus* Grade 100%

Taxa	ATCC[a]	CCUG[b]	NCTC[c]	NZRM[d]	Other[e]	Growth conditions	Length (nt)	EpsiloFsa Best Hit	nr Best Hit
H. mesocricetorum	**700932[T]**	45420[T]				MAC at 37°C on CBA	1443		*H. mesocricetorum* Grade 98.8% (0 SNP, 1407 nt of 1420 nt partial sequence)
H. muridarum	49282[T]	29262[T]	12714[T]		**LMG 13646[T]**	MAC at 37°C on CBA	1447	*H. muridarum* Grade 100%	*H. muridarum* Grade 99.9% (1 SNP, 1444 nt of 1478 nt partial sequence)
H. mustelae	43772[T]	**25715[T]**	12198[T]		12198[T]	MAC at 37°C on CBA	1385	*H. mustelae* Grade 100%	*H. mustelae* Grade 100%
H. pametensis	51478[T]	**29255[T]**	12887[T]			MAC at 37°C on CBA	1434	*H. pametensis* Grade 100%	*H. pametensis* Grade 100%
H. pullorum	51801[T]	**33837[T]**	12824[T]			MAC at 37°C on CBA	1394	*H. pullorum* Grade 99.6% (8 SNP, 1394 nt of complete gene)	*H. pullorum* Grade 100%
H. pylori	**49503**				60190	MicroAero[i] at 37°C on CBA	1457	*H. pylori* Grade 99.9% (2 SNP, 1457 nt of complete gene)	*H. pylori* Grade 99.9% (2 SNP, 1457 nt of complete gene)
H. pylori	**51932**				Tx30a	MicroAero at 37°C on CBA	1454	*H. pylori* Grade 99.9% (2 SNP, 1454 nt of complete gene)	*H. pylori* Grade 99.9% (3 SNP, 1454 nt of complete gene)
H. rodentium	**700285[T]**					MAC at 37°C on CBA			
H. salomonis		37845[T]			**Inkinen[T]**	MAC at 37°C on CBA			
'H. winghamensis'	**BAA-430[T]**					MAC at 37°C on CBA	1449	*H. winghamensis* Grade 100%	*H. winghamensis* Grade 99.8% (0 SNP, 1443 nt of 1449 nt partial sequence)
Klebsiella pneumoniae	13883[T]	225[T]	9633[T]	**482[T]**		Aerobic at 37°C on CBA	1461		*K. pneumoniae* Grade 100%

Taxa	ATCC[a]	CCUG[b]	NCTC[c]	NZRM[d]	Other[e]	Growth conditions	Length (nt)	EpsiloFsa Best Hit	nr Best Hit
Listeria monocytogenes	35152		7973	**44**		Aerobic at 37°C on CBA	1456		*L. monocytogenes* Grade 100%
Pseudomonas aeruginosa	25668	1423	10662	**981**		Aerobic at 37°C on TSA	1486		*P. aeruginosa* Grade 100%
Salmonella Typhimurium				**3970**		Aerobic at 37°C on TSA	1420		*Salmonella enterica* (various serovars incl Typhimurium, Thompson & Paratyphi B) Grade 100%
Shigella sonnei			8220	**86**		Aerobic at 37°C on TSA	1480		*E. coli/S. flexneri* Grade 99.9% (3 N & 1 Y in query, 1479 nt of complete gene), *S. sonnei* Grade 99.9% (3 N & 1 Y in query, 1 deletion, 1479 nt of complete gene), the sequence generated is not sufficient quality to differentiate these species
Staphylococcus aureus	25923			**917**		Aerobic at 37°C on CBA	1442		*S. aureus* Grade 100%
Vibrio parahaemolyticus	43996		10884	**820**		Aerobic at 37°C on CBA	631 & 445		*V. parahaemolyticus/V. natriegens/V. neocaledonicus/V. alginolyticus/V. antiquarius* Grade 99.9% (0 SNP, 630 nt of 1471 nt partial sequence) & *V. parahaemolyticus* Grade 100% (0 SNP, 445 nt of 891nt partial sequence)
Vibrio vulnificus	27562^T	13448^T		**2506^T**		Aerobic at 37°C on CBA	781 & 378		*V. vulnificus* Grade 99.9% (1 deletion, 780 nt nt of 1468 nt partial sequence) & *V. vulnificus* Grade 100%
W. succinogenes	29543^T	**13145^T**	11488^T			MAC at 37°C on CBA	1448	*C. concisus* Grade 100%	*C. concisus* Grade 100%

Taxa	ATCC[a]	CCUG[b]	NCTC[c]	NZRM[d]	Other[e]	Growth conditions	Length (nt)	EpsiloFsa Best Hit	nr Best Hit
Yersinia enterocolitica		33553	10460	**767**		Aerobic at 37°C on TSA	1511		*Y. enterocolitica* Grade 99.9% (4 SNP including first 2 nt and last nt, 1508 nt of complete gene)
Yersinia pseudotuberculosis	29833^T	5855^T	10275^T	**768^T**		Aerobic at 37°C on TSA	1507		*Y. pseudotuberculosis/Y. pestis* Grade 100%

[a] American Type Culture Collection, USA; [b] Culture Collection, University of Göteborg, Sweden; [c] National Collection of Type Cultures, Public Health England, UK; [d] New Zealand Reference Culture Collection, Medical Section, New Zealand; [e] CMB, ESR Christchurch Molecular Biology Laboratory; CPH, ESR Christchurch Public Health Laboratory; DSM, Leibniz Institute DSMZ – German Collection of Microorganisms and Cell Cultures, Germany; LMG, Laboratory of Microbiology, Ghent University, Belgium; [f] Columbia Sheep Blood Agar; [g] MAC-VA 500 cabinet (Don Whitely Scientific) with an atmosphere of 10% CO_2, 7% H_2, 3% O_2 and 80% N_2; [h] Anaero gas generation pack (Mitsubishi Gas Chemical) with an atmosphere containing 0% O_2 and 16% CO_2; [i] MicroAero gas generation pack (Mitsubishi Gas Chemical) with an atmosphere of 6-12% O_2, 5-8% CO_2; [j] Tryptic Soy Agar. Strain designations in bold represent the source of the isolate used to generate the DNA.

Table 20: Summary of Control DNA Extracts Provided by Collaborators to Evaluate the Specificity of the Epsilonproteobacteria Multiplex Ligation-Dependent Probe Amplification (MLPA) Assay

The Best Hits from both the local EpsiloFsa and NCBI nr databases were included for Epsilonproteobacteria taxa and only the NCBI nr database for other taxa. The Best Hit was the Taxon for which the highest Grade was recorded. Grade is a percentage calculated by Geneious from the query coverage, e-value and identity and aides in the identification of the longest, highest identity hits.

Taxa	ATCC[a]	CCUG[b]	NCTC[c]	NZRM[d]	Other[e]	DNA received	Length (nt)	EpsiloFsa Best Hit	nr Best Hit
A. bivalviorum					**CECT 7835^T**, RM 15224^T	Dried from MF[f]			
A. bivalviorum					CECT 7835^T, **RM 15224^T**	Dried from WM[g]	1438	*A. bivalviorum* Grade 100%	*A. bivalviorum* Grade 100%
A. cibarius		48482^T			**CECT 7203^T**, LMG 21996^T	Dried from MF			
A. cibarius					**RM 5243**	Dried from WM	1468	*A. cibarius* Grade 100%	*A. cibarius* Grade 99.9% (1 N in query, 1458 nt of 1468 nt partial sequence)
A. cloacae					**CECT 7834^T**, RM 15227^T	Dried from MF			
A. cloacae					CECT 7834^T, **RM 15227^T**	Dried from WM	1468	*A. cloacae* Grade 100%	*A. suis* Grade 98.7% (16 SNP, 1447 nt of 1454 nt partial sequence), *A. cloacae* Grade 97.8% (0 SNP, 1402 nt of 1402 nt partial sequence)
A. defluvii					CECT 7697^T, **RM 14018^T**	Dried from WM	1468	*A. defluvii* Grade 100%	*A. suis* Grade 98.7% (1 R in query, 15 SNP, 1447 nt of 1454 nt partial sequence), *A. defluvii* Grade 97.7% (1 R in query, 1402 nt of 1402 nt partial sequence)

Taxa	ATCC[a]	CCUG[b]	NCTC[c]	NZRM[d]	Other[e]	DNA received	Length (nt)	EpsiloFsa Best Hit	nr Best Hit
A. ellisii					**CECT 7837^T**, RM 15222^T	Dried from MF			
A. ellisii					CECT 7837^T, **RM 15222^T**	Dried from WM	1482	*A. ellisii* Grade 100%	*A. suis* Grade 98.2% (17 SNP, 1447 nt of 1454 nt partial sequence), *A. ellisii* Grade 97.3% (1 SNP, 1402 nt of 1402 nt partial sequence)
A. halophilus	BAA-1022^T	53805^T			**LA31B^T**, RM 5350^T	Dried from MF			
A. halophilus	BAA-1022^T	53805^T			LA31B^T, **RM 5350^T**	Dried from WM	1424	*A. halophilus* Grade 100%	*A. halophilus* Grade 99.2% (0 SNP, 1402 nt of 1402 nt partial sequence)
A. marinus					**CECT 7727^T**, RM 14021^T	Dried from MF			
A. marinus					CECT 7727^T, **RM 14021^T**	Dried from WM	1475		*A. marinus* Grade 98.1% (1 SNP, 1421 nt of 1426 nt partial sequence)
A. molluscorum					**CECT 7696^T**, RM 14015^T	Dried from MF			
A. molluscorum					CECT 7696^T, **RM 14015^T**	Dried from WM	1405	*A. molluscorum* Grade 100%	*A. molluscorum* Grade 99.1% (0 SNP, 1380 nt of 1401 nt partial sequence)
A. mytili					**CECT 7386^T**, RM 14013^T	Dried from MF			
A. mytili					CECT 7386^T, **RM 14013^T**	Dried from WM	1465	*A. mytili* Grade 100%	*A. mytili* Grade 99.4% (1 SNP, 1448 nt of 1464 nt partial sequence)
A. nitrofigilis	33309^T	15893^T	12251^T		**CECT 7204^T**, DSM 7299^T, RM 3221^T	Dried from MF			

Taxa	ATCC[a]	CCUG[b]	NCTC[c]	NZRM[d]	Other[e]	DNA received	Length (nt)	EpsiloFsa Best Hit	nr Best Hit
A. nitrofigilis	33309[T]	15893[T]	12251[T]		CECT 7204[T], DSM 7299[T], **RM 3221[T]**	Dried from WM	1463	*A. nitrofigilis* Grade 99.9% (2 N in query sequence, 1463 nt of complete gene)	*A. nitrofigilis* Grade 99.9% (2 N in query, 1463 nt of 1505 nt partial sequence)
A. suis					CECT 7833[T], **RM 15228[T]**	Dried from WM	1466	*A. suis* Grade 100%	*A. suis* Grade 99.8% (0 SNP, 1445 nt of 1454 nt partial sequence)
A. thereius		56902[T]			**LMG 24486[T]**, RM 5348[T]	Dried from MF			
A. thereius		56902[T]			LMG 24486[T], **RM 5348[T]**	Dried from WM	1460	*A. thereius* Grade 100%	*A. thereius* Grade 99.2% (15 SNP, 1 N in hit, 1452 nt of 1483 nt partial sequence)
A. trophiarum		59229[T]			**LMG 25534[T]**, RM 12658[T]	Dried from MF			
A. trophiarum		59229[T]			LMG 25534[T], **RM 12658[T]**	Dried from WM	1485	*A. trophiarum* Grade 100%	*A. thereius* Grade 98.6% (24 SNP, 1 N in hit, 1468 nt of 1483 nt partial sequence). *A. trophiarum* Grade 94.0% (1 SNP, 1308 nt of 1309 nt partial sequence)
A. venerupis					**F67-11[T]**, RM 16046[T]	Dried from MF			
A. venerupis					F67-11[T], **RM 16046[T]**	Dried from WM	1458	*A. venerupis* Grade 100%	*A. suis* Grade 98.6% (31 SNP, 1 insertion , 1449 nt of 1453 nt partial sequence), *A. venerupis* Grade 98.0% (1 SNP, 1401 nt of 1401 nt partial sequence)

Taxa	ATCC[a]	CCUG[b]	NCTC[c]	NZRM[d]	Other[e]	DNA received	Length (nt)	EpsiloFsa Best Hit	nr Best Hit
C. avium		56292[T]			**RM 8639[T]**	Dried from WM	1427	*C. avium* Grade 100%	*C. avium* Grade 99.8% (0 SNP, 1422 nt of 1435 nt partial sequence)
C. canadensis					L267, **RM 9173**	Dried from WM	1438	*C. canadensis* Grade 100%	*C. canadensis* Grade 97.1% (0 SNP, 1354 nt of 1354 nt partial sequence)
C. cuniculorum		56289[T]			DSM 23162[T], **RM 8641[T]**	Dried from WM	1454	*C. cuniculorum* Grade 100%	*C. cuniculorum* Grade 98.5% (0 SNP, 1411 nt of 1414 nt partial sequence)
C. helveticus					**ACP123b**	Liquid from KB[h]	1436	*C. helveticus* Grade 100%	*C. helveticus* Grade 99.5% (1 SNP, 1 deletion, 1425 nt of 1439 nt partial sequence)
C. helveticus					**ACP141a**	Liquid from KB	1313	*C. helveticus* Grade 99.7% (7 SNP, 1313 nt of complete gene)	*C. peloridis* Grade 98.6% (36 SNP, 1312 nt of 1493 nt partial sequence), *C. helveticus* Grade 98.6% (7 SNP, 1 deletion, 1282 nt of 1439 nt partial sequence)
C. helveticus					**ACP175a**	Liquid from KB	1238	*C. helveticus* Grade 100%	*C. helveticus* Grade 100%
C. insulaenigrae		48653[T]	12927[T]		**RM 5435[T]**	Dried from WM	1419	*C. insulaenigrae* Grade 100%	*C. insulaenigrae* Grade 100%
C. lari subsp. *concheus*		55786[T]			**RM 14091[T]**	Dried from WM	1448	*C. lari* subsp. *concheus*/UPTC Grade 100%	*C. lari* subsp. *concheus*/UPTC Grade 100%

Taxa	ATCC[a]	CCUG[b]	NCTC[c]	NZRM[d]	Other[e]	DNA received	Length (nt)	EpsiloFsa Best Hit	nr Best Hit
C. sputorum biovar *sputorum*	35980[T]	9728[T]	11528[T]		**CIP 103749[T]**	Liquid from FM[i]	1427	*C. lari* Grade 88.6% (Pairwise Identity & Identical Sites 77.1%, 1441 nt). N.B. No *C. sputorum* hits with this database. When aligned against a 16S rRNA extracted from RM4121 Pairwise Identity & Identical Sites 71.5%	*Staphylococcus capitis* Grade 100%
C. sputorum biovar *sputorum*					**RM 3237**	Dried from WM	Sequence too poor to generate sequence for BLAST searching		
C. upsaliensis					ACP170b[j]	Liquid from AM			
C. upsaliensis					L395	Liquid from OV[k]			
C. volucris		57498[T]			**RM 9726[T]**	Dried from WM	1463	*C. volucris* Grade 100%	*C. volucris* Grade 100%
H. bilis	51630[T]	38995[T]			**CIP 104752[T]**, RM 3240[T]	Liquid from FM	1612	*H. bilis* Grade 100%	*H. bilis* Grade 100%
H. bilis	51630[T]	38995[T]			CIP 104752[T], **RM 3240[T]**	Dried from WM	1612	*H. bilis* Grade 100%	*H. bilis* Grade 100%

Taxa	ATCC[a]	CCUG[b]	NCTC[c]	NZRM[d]	Other[e]	DNA received	Length (nt)	EpsiloFsa Best Hit	nr Best Hit
H. bizzozeronii		35545^T			**CIP 105233^T**	Liquid from FM	1464	*Caminibacter mediatlanticus* Grade 88.5% (Pairwise Identity & Identical Sites 77.1%, 1483 nt). *H. bizzozeronii* Grade 85.4% (Pairwise Identity 76.7%, Identical Sites 76.6%, 1395 nt. 16S rRNA sequence extracted from the genome for this strain (CCUG 35545) Pairwise Identity 72.7% & Identical Sites 71.1%	*Corynebacterium tuberculostearicum* Grade 99.9%
H. cholecystus	700242^T				**CIP 105596^T**	Liquid from FM	1386		*H. cholecystus* Grade 100%
H. ganmani		43526^T			**CIP 106846^T**	Liquid from FM	1381		*H. ganmani* Grade 100%
H. typhlonius	BAA-367^T	48335^T			**CIP 107729^T**, MIT 98-6810^T	Liquid from FM	1605	*H. typhlonius* Grade 100%	*H. typhlonius* Grade 100%

[a] American Type Culture Collection, USA; [b] Culture Collection, University of Göteborg, Sweden; [c] National Collection of Type Cultures, Public Health England, UK; [d] New Zealand Reference Culture Collection, Medical Section, New Zealand; [e] CECT, Spanish Type Culture Collection, Universitat de València, Spain; CIP, Collection of the Institut Pasteur, France; LMG, Laboratory of Microbiology, Ghent University, Belgium. Strain designations in bold represent the source of the isolate used to generate the DNA. [f] Dr Maria Figueras (Universitat Rovira i Virgili, Spain), [g] Dr William Miller (United States Department of Agriculture, USA), [h] Krunoslav Bojanic (Massey University), [i] Dr Francis Megraud (University of Bordeaux, France), [j] Anne Midwinter (Massey University), [k] Prof Olivier Vandenberg (St. Pierre University Hospital, Belgium). Strain designations in bold represent the source of the isolate used to generate the DNA.

16S rRNA sequencing was undertaken on 95 of the extracts. The sequence obtained for DNA extracts from two *C. sputorum* biovar *sputorum* strains (CCUG 9728^T and RM 3237) was not of sufficient quality to obtain an identification in spite of multiple attempts at 16S rRNA sequencing. The sequence lengths for the remaining 93 extracts ranged from 232 nt for the *C. sputorum* biovar *sputorum* strain CCUG 11289 to 1786 nt for the *H. fennelliae* strain CCUG 18820^T. For the majority (88, 92.6%) of DNA extracts a single sequence was obtained. For seven DNA extracts two sequences, both greater than 350 nt, were generated and aided in confirming the identification for the DNA. The expect value, a measure of how many sequences matching the query could be expected from the database purely by chance (Sansom 2000), was 0 for all queries except the CCUG 11289 results which were $6.56x10^{-114}$ for the EpsiloFsa database and $5.17x10^{-112}$ for the nr database, which is to be expected when only a short sequence was generated.

The 16S rRNA sequence-based identification results for 12 of the DNA extracts were unexpected. 16S rRNA sequencing was attempted on three DNA extracts from two *C. sputorum* biovar *sputorum* strains. For two of the DNA extracts (CCUG 9728^T and RM 3237), the sequence generated was too poor to be able to generate a consensus sequence for BLAST searching. To investigate the inability to generate 16S rRNA sequence for these two DNA extracts, the primers used for sequencing were aligned to the 16S rRNA sequences extracted from the four *C. sputorum* genomes used in this study. As shown in Figure 37, three mismatches were found with the 10 primers. There are mismatches at nt 3 of 16F945, nt 11 of 16R1389 and nt 15 of 16R1541. The 1427 nt sequence obtained for the remaining *C. sputorum* biovar *sputorum* strain DNA extract (CIP 103749^T) showed 100% identity to *Staphylococcus capitis*. The 1464 nt sequence obtained for DNA extract from the *H. bizzozeronii* strain CIP 105233^T showed greatest similarity to *Corynebacterium tuberculostearicum*. The quantity of DNA for this strain was 3.0 ng/µL upon receipt so it is possible that the DNA actually amplified for this extract was a contaminant. The 1487 nt sequence from the *C. rectus* strain CCUG 20446^T showed greatest similarity to *C. ureolyticus*. An alternative DNA extract for this strain,

with the designation ATCC 33238^{T}, was available but there was insufficient volume to conduct 16S rRNA sequencing on this extract. The 1448 nt sequence from the *W. succinogenes* strain CCUG 13145^{T} showed greatest similarity to *C. concisus*. Attempts to produce a new DNA extract for this strain were unsuccessful because the frozen culture had become non-viable.

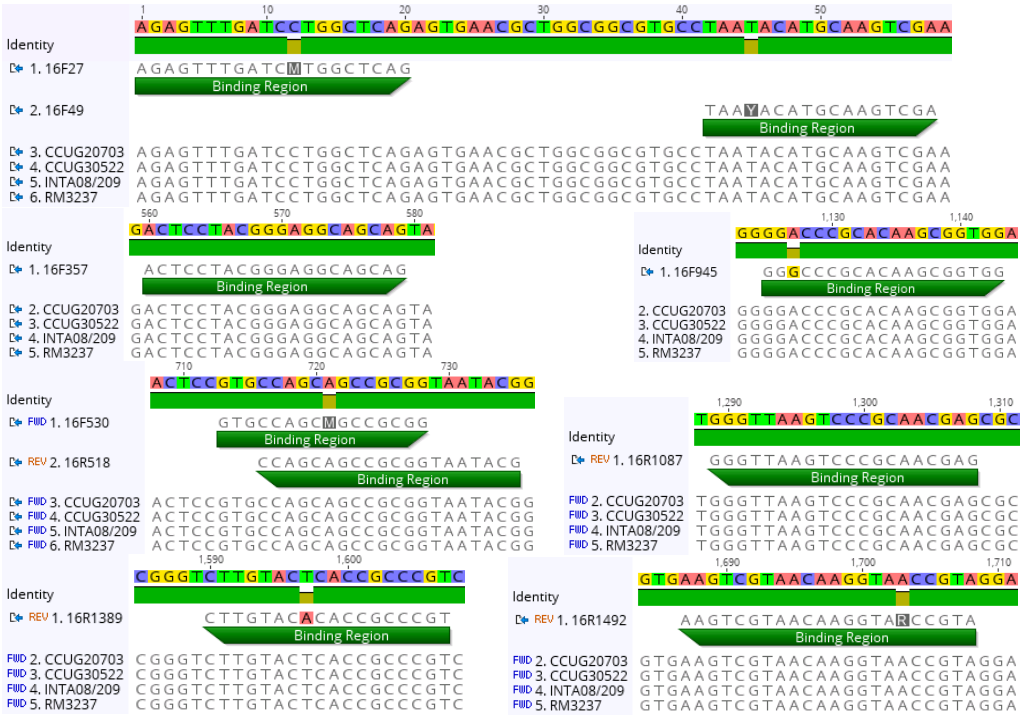


Figure 37: Alignment of the 16S rRNA Gene Sequencing Primers with 16S rRNA Gene Sequences Extracted from Four *C. sputorum* Genomes

The sequences obtained for the DNA extracts of the *Shigella sonnei* strain NZRM 86, *Bacillus cereus* strain NZRM 5 and *Yersinia pseudotuberculosis* strain NZRM 768^{T} resulted in uncertain taxonomic identifications. The 1480 nt *Shigella sonnei* sequence contained 4 ambiguous nucleotides and was otherwise indistinguishable from *Escherichia coli* and *Shigella flexneri* as well as *Shigella sonnei* sequences. Similarly the 1442 nt *Bacillus cereus* sequence from the strain NZRM 5 was indistinguishable from sequences from *Bacillus anthracis, Bacillus toyonensis, Bacillus thuringiensis* and *Bacillus cereus*, and the 1507 nt sequence from the *Yersinia pseudotuberculosis* strain NZRM 768^{T} was indistinguishable from sequences from both *Yersinia pestis* and *Yersinia pseudotuberculosis*. The 1449 nt and 1437 nt sequences obtained from the two *C. fetus* subspecies were both indistinguishable from sequences from both subspecies. The 232 nt sequence obtained for the *C. sputorum* subsp. *bubulus* strain CCUG

11289^T was indistinguishable from the sequences from both *C. sputorum* subspecies. The remaining 83 (87%) DNA extracts returned the expected identifications although the expected taxa did not have the highest Grade for the sequences from nine strains (*A. cloacae* RM15227^T, *A. cryaerophilus* CCUG 17801^T, *A. defluvii* RM 14018^T, *A. ellisii* RM 15222^T, *A. trophiarum* RM 12658^T, *A. venerupis* RM 16046^T, *C. cuniculorum* CCUG 56289^T, *C. helveticus* ACP141a and *C. showae* CCUG 30254^T), generally because the hit sequence of the expected taxa was shorter than the sequences of both the query and non-target taxa. In addition, the maximum number of hits returned for the nr search had to be increased from 100 to 1000 in order to return the target taxa for sequences generated for the *C. jejuni subsp. doylei* strain CCUG24567^T and *A. trophiarum* strain RM 12658^T.

Table 21 summarises the MLPA results for the DNA extracts with discordant 16S rRNA sequence-based identifications. The DNA extract from the *C. rectus* strain CCUG 20446^T was positive for the Cureolyticus MLPA probe confirming the 16S rRNA sequencing result for this extract. Similarly, the DNA extract from the *W. succinogenes* strain CCUG13145^T was positive for the Cconcisus and CconcicusGS1 probes confirming the 16S rRNA sequencing result. The results for these two DNA extracts were excluded from further analysis. The Epsilonproteobacteria MLPA results for the *C. sputorum* and *H. bizzozeronii* DNA extracts were concordant with expectations but since neither of these taxa are included in the possible targets for the Epsilonproteobacteria MLPA assay (Table 8) these DNA extracts were also excluded in the evaluation of the assay.

Table 21: Epsilonproteobacteria Multiplex Ligation-dependent Probe Amplification (MLPA) Results for DNA Extracts with Discordant Identification

Taxa	Strain	Concentration	Rep	MLPA probes detected[a]
C. rectus	CCUG 20446^T	20 ng/µL		Cureolyticus
C. sputorum	CCUG 11289	20 ng/µL	a	Campylobacter
biovar *sputorum*			b	Campylobacter
	CCUG 9728^T	20 ng/µL	a	Campylobacter
			b	Campylobacter
	CIP 103749^T	18.7 ng/µL		Campylobacter
	RM 3237	17.9 ng/µL		Campylobacter
H. bizzozeronii	CIP 105233^T	3 ng/µL	a	None
			b	None
W. succinogenes	CCUG 13145^T	20 ng/µL		Cconcisus + CconcisusGS1

[a] Probes in bold were not expected based on the taxa of the DNA extracts

Discussion

Only two of the four *C. sputorum* DNA extracts sent for 16S rRNA sequencing generated useable sequence. The sequence for CIP 103749^T was 1427 nt was identical to a *Staphylococcus capitis* 16S rRNA gene sequence, except for the N in the query sequence and the sequence for CCUG 11289 was 232 nt long and had over 99% identity to *C. sputorum* 16S rRNA sequences in both the local EpsiloFsa and NCBI nr databases. All four *C. sputorum* DNA extracts produced a Campylobacter peak with the Epsilonproteobacteria MLPA, as would be expected for *C. sputorum* DNA. It is possible that the CIP 103749^T DNA had become contaminated with *Staphylococcus capitis* DNA but with enough *C. sputorum* DNA remaining to anneal to the Campylobacter MLPA probes. *C. sputorum* has been shown to have an intervening sequence within the 16S rRNA gene (Etoh, Yamamoto, and Goto 1998, Tazumi et al. 2010) but there are no reports of difficulty generating significant sequence information for this gene in this species. Alignment of the 16S rRNA sequencing primers with 16S rRNA sequences extracted from the four *C. sputorum* genomes identified three mismatches with the 10 primers. The mismatch at nt 15 of 16R1541, which is only 6 nucleotides from the 3' end of the reverse primer

used to amplify the gene for sequencing, is likely the reason why it is not always possible to generate 16S rRNA sequence for this species.

The 16S rRNA sequences for the *C. rectus* CCUG 20446^T, *W. succinogenes* CCUG 13145^T and *H. bizzozeronii* CIP 105233^T showed highest similarity to *C. ureolyticus, C. concisus* and *Corynebacterium tuberculostearicum*, respectively. The Epsilonproteobacteria MLPA results for CCUG 20446^T and CCUG 13145^T support the 16S rRNA results and suggest that these DNA extracts, or the cultures they were produced from, had become contaminated. The MLPA results for these two extracts were ignored during the evaluation of the performance of the MLPA assay. CIP 105233^T was negative for all probes in the Epsilonproteobacteria MLPA assay. Given that this species was not a target of any MLPA probe, this result provides no additional information for establishing whether the CIP 105223^T extract contains any DNA from this species. Nor does the result influence the performance evaluation of the MLPA assay.

The BLAST searches for the 16S rRNA sequences generated for *Bacillus cereus, Shigella sonnei* and *Yersinia pseudotuberculosis* obtained uncertain identifications due to sequences for multiple species in the NCBI nr database having equal similarity to the query sequence. These results are concordant with previous publications for these species (Ibrahim et al. 1993, Jimenez et al. 2013, Kotetishvili et al. 2005, Lukjancenko, Wassenaar, and Ussery 2010).

16S rRNA is a useful method for confirming the identification of DNA extracts being used to evaluate molecular assays. It is important that the BLAST results are carefully reviewed to ensure the correct identification, which may not produce the highest ranking hit, is obtained. In addition, the discriminatory power of this gene for each target genus should be considered when interpreting the 16S rRNA sequence information.

Appendix IV: Specificity of the Epsilonproteobacteria MLPA Assay

Taxa	Strain	Concentration	Rep	MLPA probes detected[a]	Note
A. bivalviorum	CECT 7835^T	low		Arcobacter	resuspended in dH_2O
	RM 15224^T	21 ng/µL	a	Arcobacter	resuspended in dH_2O
			b	Arcobacter	resuspended in dH_2O
A. butzleri	A758	20 ng/µL		Arcobacter + Abutzleri	eluted in 10 mM Tris, 0.5 mM EDTA, pH 9.0
	BRUG0093	20 ng/µL		Arcobacter + Abutzleri	diluted in low EDTA-TE Buffer
	BRUG0198	20 ng/µL		Arcobacter + Abutzleri	diluted in low EDTA-TE Buffer
	BRUG0318	20 ng/µL		Arcobacter + Abutzleri	diluted in low EDTA-TE Buffer
	CCUG 30485^T	20 ng/µL	a	Arcobacter + Abutzleri	
			b	Arcobacter + Abutzleri	
		2 ng/µL		Arcobacter + Abutzleri	diluted in low EDTA-TE Buffer
A. cibarius	CECT 7203^T	low		Arcobacter	resuspended in dH_2O
	RM 5243	9.1 ng/µL	a	Arcobacter	resuspended in dH_2O
			b	Arcobacter	resuspended in dH_2O
		0.91 ng/µL		Arcobacter	diluted in low EDTA-TE Buffer
A. cloacae	CECT 7834^T	low		Arcobacter	resuspended in dH_2O
	RM 15227^T	7.9 ng/µL	a	Arcobacter	resuspended in dH_2O
			b	Arcobacter	resuspended in dH_2O
		0.79 ng/µL		Arcobacter	diluted in low EDTA-TE Buffer

Taxa	Strain	Concentration	Rep	MLPA probes detected[a]	Note
A. cryaerophilus	CCUG 17801^T	20 ng/µL	a	Arcobacter + Acryaerophilus	
			b	Arcobacter + Acryaerophilus	
		2 ng/µL		Arcobacter + Acryaerophilus	diluted in low EDTA-TE Buffer
A. defluvii	RM 14018^T	16.2 ng/µL	a	Arcobacter	resuspended in dH_2O
			b	Arcobacter	resuspended in dH_2O
		1.62 ng/µL		Arcobacter	diluted in low EDTA-TE Buffer
A. ellisii	CECT 7837^T	low		Arcobacter	resuspended in dH_2O
	RM 15222^T	23 ng/µL	a	Arcobacter	resuspended in dH_2O
			b	Arcobacter	resuspended in dH_2O
		2.3 ng/µL		Arcobacter	diluted in low EDTA-TE Buffer
A. halophilus	LA31B^T	low		Arcobacter	resuspended in dH_2O
	RM 5350^T	20 ng/µL	a	Arcobacter	resuspended in dH_2O
			b	Arcobacter	resuspended in dH_2O
		2 ng/µL		Arcobacter	diluted in low EDTA-TE Buffer
A. marinus	CECT 7727^T	low		Arcobacter	resuspended in dH_2O
	RM 14021^T	15.7 ng/µL	a	Arcobacter	resuspended in dH_2O
			b	Arcobacter	resuspended in dH_2O
		1.57 ng/µL		Arcobacter	diluted in low EDTA-TE Buffer
A. molluscorum	CECT 7696^T	low		Arcobacter	resuspended in dH_2O

Taxa	Strain	Concentration	Rep	MLPA probes detected[a]	Note
A. molluscorum	RM 14015^T	19.4 ng/μL	a	Arcobacter	resuspended in dH_2O
			b	Arcobacter	resuspended in dH_2O
		1.94 ng/μL		Arcobacter	diluted in low EDTA-TE Buffer
A. mytili	CECT 7386^T	low		Arcobacter	resuspended in dH_2O
	RM 14013^T	13.9 ng/μL	a	Arcobacter	resuspended in dH_2O
			b	Arcobacter	resuspended in dH_2O
		1.39 ng/μL		Arcobacter	diluted in low EDTA-TE Buffer
A. nitrofigilis	CECT 7204^T	low		Arcobacter	resuspended in dH_2O
	RM 3221^T	20 ng/μL	a	Arcobacter	resuspended in dH_2O
			b	Arcobacter	resuspended in dH_2O
		2 ng/μL		Arcobacter	diluted in low EDTA-TE Buffer
A. skirrowii	CCUG 10374^T	20 ng/μL	a	Arcobacter	
			b	Arcobacter	
		2 ng/μL		Arcobacter	diluted in low EDTA-TE Buffer
A. suis	RM 15228^T	10 ng/μL	a	Arcobacter	resuspended in dH_2O
			b	Arcobacter	resuspended in dH_2O
		1 ng/μL		Arcobacter	diluted in low EDTA-TE Buffer
A. thereius	LMG 24486^T	low		Arcobacter	resuspended in dH_2O
	RM 5348^T	18.4 ng/μL	a	Arcobacter	resuspended in dH_2O
			b	Arcobacter	resuspended in dH_2O
A. trophiarum	LMG 25534^T	low		Arcobacter	resuspended in dH_2O

Taxa	Strain	Concentration	Rep	MLPA probes detected[a]	Note
A. trophiarum	RM 12658^T	9.6 ng/µL	a	Arcobacter	resuspended in dH_2O
			b	Arcobacter	resuspended in dH_2O
		0.96 ng/µL		Arcobacter	diluted in low EDTA-TE Buffer
A. venerupis	F67-11^T	low		Arcobacter	resuspended in dH_2O
	RM 16046^T	18.9 ng/µL	a	*Arcobacter*	resuspended in dH_2O
			b	Arcobacter	resuspended in dH_2O
Bacillus cereus	NZRM 5	20 ng/µL		None	diluted in low EDTA-TE Buffer
C. avium	RM 8639^T	24 ng/µL	a	Campylobacter + Cavium	resuspended in dH_2O
			b	Campylobacter + Cavium	resuspended in dH_2O
C. canadensis	CCUG 54429^T	20 ng/µL	a	Ccanadensis	Campylobacter not detected
			b	Ccanadensis	Campylobacter not detected
		2 ng/µL		Campylobacter + Ccanadensis	diluted in low EDTA-TE Buffer
	RM 9173	5.1 ng/µL		Campylobacter + Ccanadensis	resuspended in dH_2O
C. coli	CCUG 11283^T	20 ng/µL	a	Campylobacter + Ccoli	
			b	Campylobacter + Ccoli + Ccuniculorum	
		2 ng/µL		Campylobacter + Ccoli	diluted in low EDTA-TE Buffer
	CMB06690	20 ng/µL		Ccoli	prepared in low EDTA-TE Buffer, Campylobacter not detected
	CMB091229	20 ng/µL		Campylobacter + Ccoli	prepared in low EDTA-TE Buffer
	CPH0311750	20 ng/µL		Ccoli	prepared in low EDTA-TE Buffer, Campylobacter not detected
	RM 2228	20 ng/µL		Ccoli	prepared in low EDTA-TE Buffer, Campylobacter not detected

Taxa	Strain	Concentration	Rep	MLPA probes detected[a]	Note
C. concisus GS1	CCUG 13144^T	20 ng/µL	a	Cconcisus + CconcisusGS1	Campylobacter not detected
			b	Cconcisus	Campylobacter + CconcicusGS1 not detected
		2 ng/µL		Campylobacter + Cconcisus + CconcisusGS1	diluted in low EDTA-TE Buffer
	Lasto115.99	20 ng/µL	a	Campylobacter + Cconcisus + CconcisusGS1	
			b	Cconcisus + CconcisusGS1	Campylobacter not detected
		2 ng/µL		Campylobacter + Cconcisus + CconcisusGS1	diluted in low EDTA-TE Buffer
	Lasto220.96	20 ng/µL	a	Cconcisus + CconcisusGS1	Campylobacter not detected
			b	Cconcisus + *CconcisusGS1*	Campylobacter not detected
		2 ng/µL		Campylobacter + Cconcisus + CconcisusGS1	diluted in low EDTA-TE Buffer
	Lasto24.99	20 ng/µL	a	Cconcisus	Campylobacter + CconcisusGS1 not detected
			b	Cconcisus	Campylobacter + CconcisusGS1 not detected
		2 ng/µL		Campylobacter + Cconcisus + CconcisusGS1 + CconcisusGS2	diluted in low EDTA-TE Buffer

Taxa	Strain	Concentration	Rep	MLPA probes detected[a]	Note
C. concisus GS1	Lasto28.99	20 ng/µL	a	Campylobacter + Cconcisus	CconcisusGS1 not detected
			b	Cconcisus	Campylobacter + CconcisusGS1 not detected
		2 ng/µL		Campylobacter + Cconcisus + CconcisusGS1	diluted in low EDTA-TE Buffer
	Lasto389.96	20 ng/µL	a	Campylobacter + Cconcisus + CconcisusGS1	
			b	Campylobacter + Cconcisus + CconcisusGS1	
	Lasto393.96	2 ng/µL		Campylobacter + Cconcisus + CconcisusGS1	diluted in low EDTA-TE Buffer
	Lasto61.99	20 ng/µL	a	Cconcisus + CconcisusGS1	Campylobacter not detected
			b	Campylobacter + Cconcisus + CconcisusGS1	
		2 ng/µL		Campylobacter + Cconcisus + CconcisusGS1	diluted in low EDTA-TE Buffer
	Lasto64.99	19 ng/µL	a	Campylobacter + Cconcisus + CconcisusGS1	eluted in 10 mM Tris, 0.5 mM EDTA, pH 9.0
			b	Campylobacter + Cconcisus + CconcisusGS1	eluted in 10 mM Tris, 0.5 mM EDTA, pH 9.0
		2 ng/µL		Campylobacter + Cconcisus + CconcisusGS1	diluted in low EDTA-TE Buffer

Taxa	Strain	Concentration	Rep	MLPA probes detected[a]	Note
C. concisus GS2	CCUG 19995	20 ng/µL	a	Campylobacter + Cconcisus + CconcisusGS2	
			b	Cconcisus	Campylobacter + CconcisusGS2 not detected
		2 ng/µL		Campylobacter + Cconcisus + CconcisusGS2	diluted in low EDTA-TE Buffer
	Lasto104.93	20 ng/µL	a	None	No Q-fragments
			b	Campylobacter + Cconcisus	CconcisusGS2 not detected
			c	Campylobacter + Cconcisus	CconcisusGS2 not detected
		2 ng/µL		Campylobacter + Cconcisus + CconcisusGS2+ CconcisusGS1	diluted in low EDTA-TE Buffer
C. concisus GS2	Lasto113.99	20 ng/µL	a	Cconcisus + CconcisusGS2	Campylobacter not detected
			b	Campylobacter + Cconcisus + CconcisusGS2	
		2 ng/µL		Campylobacter + Cconcisus + CconcisusGS2	diluted in low EDTA-TE Buffer
	Lasto131.99	20 ng/µL	a	None	No Q-fragments
			b	Cconcisus	Campylobacter + CconcisusGS2 not detected
			c	Campylobacter + Cconcisus + CconcisusGS2	
		2 ng/µL		Campylobacter + Cconcisus + CconcisusGS2	diluted in low EDTA-TE Buffer

Taxa	Strain	Concentration	Rep	MLPA probes detected[a]	Note
C. concisus GS2	Lasto135.99	20 ng/µL	a	Campylobacter + Cconcisus + CconcisusGS2	
			b	Campylobacter + Cconcisus + CconcisusGS2	
		2 ng/µL		Campylobacter + Cconcisus + CconcisusGS2	diluted in low EDTA-TE Buffer
	Lasto140.99	20 ng/µL	a	Cconcisus	Campylobacter + CconcisusGS2 not detected
			b	Cconcisus	Campylobacter + CconcisusGS2 not detected
		2 ng/µL		Campylobacter + Cconcisus + CconcisusGS2	diluted in low EDTA-TE Buffer
C. concisus GS2	Lasto275.95	20 ng/µL	a	Campylobacter + Cconcisus + CconcisusGS2	
			b	Campylobacter + Cconcisus + CconcisusGS2	
	Lasto316.98	20 ng/µL	a	Campylobacter + Cconcisus + CconcisusGS2	
			b	Cconcisus + CconcisusGS2	Campylobacter not detected
		2 ng/µL		Campylobacter + Cconcisus + CconcisusGS2	diluted in low EDTA-TE Buffer
	Lasto377.96	20 ng/µL	a	Campylobacter + Cconcisus + CconcisusGS2	
			b	Cconcisus + CconcisusGS2	Campylobacter not detected
		2 ng/µL		Campylobacter + Cconcisus + CconcisusGS2	diluted in low EDTA-TE Buffer

Taxa	Strain	Concentration	Rep	MLPA probes detected[a]	Note
C. cuniculorum	CCUG 56289^T	20 ng/µL	a	Ccuniculorum	Campylobacter not detected
			b	Ccuniculorum	Campylobacter not detected
	RM 8641^T	16.3 ng/µL		Campylobacter + Ccuniculorum	resuspended in dH_2O
		1.63 ng/µL		Campylobacter + Ccuniculorum	diluted in low EDTA-TE Buffer
C. curvus	CCUG 13146^T	20 ng/µL	a	Campylobacter	
			b	Campylobacter	
C. fetus subsp. *fetus*	CCUG 6823^T	20 ng/µL	a	Campylobacter	
			b	Campylobacter	
C. fetus subsp. *venerealis*	CCUG 538^T	20 ng/µL	a	Campylobacter	
			b	Campylobacter	
		2 ng/µL		Campylobacter	diluted in low EDTA-TE Buffer
C. gracilis	CCUG 27720^T	5.1 ng/µL	a	Campylobacter	prepared in low EDTA-TE buffer
			b	Campylobacter	prepared in low EDTA-TE buffer
		0.51 ng/µL		Campylobacter	diluted in low EDTA-TE Buffer
C. helveticus	ACP123b	20 ng/µL		Campylobacter + Chelveticus	diluted in low EDTA-TE Buffer
	ACP141a	20 ng/µL		Campylobacter + Chelveticus	diluted in low EDTA-TE Buffer
	ACP175a	20 ng/µL		Campylobacter + Chelveticus	diluted in low EDTA-TE Buffer
	CCUG 30682^T	20 ng/µL	a	Chelveticus	Campylobacter not detected
			b	Chelveticus	Campylobacter not detected

Taxa	Strain	Concentration	Rep	MLPA probes detected[a]	Note
C. hominis	CH001^T	20 ng/µL	a	Campylobacter	
			b	Campylobacter	
		2 ng/µL		Campylobacter	diluted in low EDTA-TE Buffer
C. hyointestinalis subsp. *hyointestinalis*	CCUG 14169^T	20 ng/µL	a	Campylobacter	
			b	Campylobacter	
C. hyointestinalis subsp. *lawsonii*	CCUG 27631^T	20 ng/µL	a	Campylobacter	
			b	Campylobacter	
C. insulaenigrae	LMG 22716^T	20 ng/µL	a	Campylobacter + Cinsulaenigrae	
			b	*Campylobacter* + *Cinsulaenigrae*	
	RM 5435^T	10.1 ng/µL		Campylobacter + Cinsulaenigrae	resuspended in dH_2O
C. jejuni subsp. *doylei*	CCUG 24567^T	20 ng/µL	a	Campylobacter + Cjejuni + Cjejunidoylei	
			b	*Campylobacter* + Cjejuni + Cjejunidoylei	
		2 ng/µL		Campylobacter + Cjejuni + Cjejunidoylei	diluted in low EDTA-TE Buffer
C. jejuni subsp. *jejuni*	CCUG 11284^T	20 ng/µL	a	Campylobacter + Cjejuni	
			b	Cjejuni	Campylobacter not detected
		2 ng/µL		Campylobacter + Cjejuni	diluted in low EDTA-TE Buffer
	NCTC 11168	20 ng/µL		Cjejuni	prepared in low EDTA-TE Buffer, Campylobacter not detected
	P110B	20 ng/µL		Campylobacter + Cjejuni	prepared in low EDTA-TE Buffer

Taxa	Strain	Concentration	Rep	MLPA probes detected[a]	Note
C. jejuni subsp. *jejuni*	RM 1221	20 ng/µL		Cjejuni	prepared in low EDTA-TE Buffer, Campylobacter not detected
	RM 1864	20 ng/µL		*Cjejuni*	prepared in low EDTA-TE Buffer, Campylobacter not detected
C. lanienae	NCTC 13004^T	20 ng/µL	a	Campylobacter	
			b	Campylobacter	
C. lari subsp. *concheus*	RM 14091^T	19 ng/µL	a	Campylobacter	resuspended in dH_2O, Clariconcheus not detected
			b	Campylobacter	resuspended in dH_2O, Clariconcheus not detected
		1.9 ng/µL		Campylobacter	diluted in reduced EDTA-TE Buffer, Clariconcheus not detected
C. lari subsp. *lari*	CCUG 23947^T	20 ng/µL	a	Campylobacter + Clarilari	
			b	Campylobacter + Clarilari	
C. mucosalis	CCUG 6822^T	20 ng/µL	a	Campylobacter	
			b	Campylobacter	
		2 ng/µL		Campylobacter	diluted in low EDTA-TE Buffer
C. peloridis	R-13342^T	20 ng/µL	a	Campylobacter + Cpeloridis	
			b	*Campylobacter* + Cpeloridis	
		2 ng/µL		Campylobacter + Cpeloridis	diluted in low EDTA-TE Buffer
C. showae	CCUG 30254^T	20 ng/µL	a	Campylobacter	
			b	Campylobacter	

Taxa	Strain	Concentration	Rep	MLPA probes detected[a]	Note
C. subantarcticus	CCUG 38513^T	20 ng/µL	a	Csubantarcticus + Arcobacter	Campylobacter not detected
			b	Campylobacter + *Csubantarcticus* + Arcobacter	
			c	Csubantarcticus + Arcobacter	Campylobacter not detected
		2 ng/µL		Campylobacter + Csubantarcticus + Arcobacter	diluted in low EDTA-TE Buffer
C. upsaliensis	ACP170b	29 ng/µL		Campylobacter	Cupsaliensis not detected
	F221	20 ng/µL		Campylobacter + Cupsaliensis	diluted in low EDTA-TE Buffer
	NZRM 3675^T	20 ng/µL	a	Cupsaliensis	Campylobacter not detected
			b	Cupsaliensis	Campylobacter not detected
	L395	6.3 ng/µL		Campylobacter	Cupsaliensis not detected
		0.63 ng/µL		Campylobacter	Diluted in 10 mM Tris (pH 8.0), Cupsaliensis not detected
C. ureolyticus	CCUG 7319^T	20 ng/µL	a	Campylobacter + Cureolyticus	
			b	Cureolyticus	Campylobacter not detected
		2 ng/µL		Campylobacter + Cureolyticus	diluted in low EDTA-TE Buffer
C. volucris	RM 9726^T	4 ng/µL	a	Campylobacter + Cvolucris	resuspended in dH_2O
			b	Campylobacter + Cvolucris	resuspended in dH_2O
		0.4 ng/µL		Campylobacter + Cvolucris	diluted in low EDTA-TE Buffer
Clostridium difficile	NZRM 2390^T	20 ng/µL		Ccuniculorum	diluted in low EDTA-TE Buffer
Clostridium perfringens	NZRM 20^T	20 ng/µL		None	diluted in low EDTA-TE Buffer
Escherichia coli	ATCC 11775^T	20 ng/µL		None	diluted in low EDTA-TE Buffer

Taxa	Strain	Concentration	Rep	MLPA probes detected[a]	Note
Escherichia coli O157	NZRM 3614	20 ng/µL		None	diluted in low EDTA-TE Buffer
H. bilis	CIP 104752^T	20 ng/µL		eHelicobacter	
	RM 3240^T	6.48 ng/µL		eHelicobacter	resuspended in dH_2O
H. canadensis	CCUG 47163^T	20 ng/µL	a	eHelicobacter	
			b	eHelicobacter	
H. canis	CCUG 32756^T	20 ng/µL	a	Hcanis	eHelicobacter not detected
			b	Hcanis	eHelicobacter not detected
		2 ng/µL		eHelicobacter + Hcanis	diluted in low EDTA-TE Buffer
H. cholecystus	CIP 105596^T	20 ng/µL	a	eHelicobacter	
			b	eHelicobacter	
		2 ng/µL		eHelicobacter	diluted in low EDTA-TE Buffer
H. cinaedi	CCUG 18818^T	20 ng/µL	a	Hcinaedi	eHelicobacter not detected
			b	Hcinaedi	eHelicobacter not detected
		2 ng/µL		eHelicobacter + Hcinaedi	diluted in low EDTA-TE Buffer
H. fennelliae	CCUG 18820^T	20 ng/µL	a	Hfennelliae	eHelicobacter not detected
			b	Hfennelliae	eHelicobacter not detected
		2 ng/µL	a	eHelicobacter + Hfennelliae	diluted in low EDTA-TE Buffer
			b	eHelicobacter + Hfennelliae + Arcobacter	diluted in low EDTA-TE Buffer
H. ganmani	CIP 106846^T	9.4 ng/µL	a	eHelicobacter	
			b	eHelicobacter	

Taxa	Strain	Concentration	Rep	MLPA probes detected[a]	Note
H. hepaticus	CCUG 33637^T	20 ng/µL	a	*eHelicobacter*	
			b	eHelicobacter	
			c	eHelicobacter	
H. mesocricetorum	ATCC 700932^T	20 ng/µL	a	eHelicobacter	
			b	eHelicobacter	
H. muridarum	LMG 13646 ^{T}	20 ng/µL	a	eHelicobacter	
			b	None	
H. mustelae	CCUG 25715^T	20 ng/µL	a	eHelicobacter	
			b	eHelicobacter	
H. pametensis	CCUG 29255^T	20 ng/µL	a	None	eHelicobacter not detected
			b	None	eHelicobacter not detected
H. pullorum	CCUG 33837^T	20 ng/µL	a	None	No Q-fragments
			b	None	No Q-fragments
		2 ng/µL		eHelicobacter	diluted in low EDTA-TE Buffer, Hpullorum not detected
H. pylori	ATCC 49503	20 ng/µL	a	None	diluted in low EDTA-TE Buffer
		2 ng/µL		None	diluted in low EDTA-TE Buffer
	ATCC 51932	20 ng/µL	a	None	diluted in low EDTA-TE Buffer
		2 ng/µL		None	diluted in low EDTA-TE Buffer
H. typhlonius	CIP 107729^T	19.8 ng/µL	a	eHelicobacter	
			b	eHelicobacter	

Taxa	Strain	Concentration	Rep	MLPA probes detected[a]	Note
'H. winghamensis'	ATCC BAA-430^T	20 ng/µL	a	eHelicobacter	
			b	eHelicobacter	
Klebsiella pneumoniae	NZRM 482^T	20 ng/µL		None	diluted in low EDTA-TE Buffer
Listeria monocytogenes	NZRM 44	20 ng/µL		None	diluted in low EDTA-TE Buffer
Pseudomonas aeruginosa	NZRM 981	20 ng/µL		None	diluted in low EDTA-TE Buffer
Salmonella Typhimurium	NZRM 3970	20 ng/µL		None	diluted in low EDTA-TE Buffer
Shigella sonnei	NZRM 86	20 ng/µL		None	diluted in low EDTA-TE Buffer
Staphylococcus aureus	NZRM 917	20 ng/µL		None	diluted in low EDTA-TE Buffer
UPTC	CCUG 20707	20 ng/µL	a	ClariUPTC + Csubantarcticus	Campylobacter not detected
			b	ClariUPTC + Csubantarcticus + Arcobacter	Campylobacter not detected
			c	ClariUPTC + Csubantarcticus	Campylobacter not detected
		2 ng/µL		Campylobacter + ClariUPTC + Csubantarcticus	diluted in low EDTA-TE Buffer
Vibrio parahaemolyticus	NZRM 820	20 ng/µL		None	diluted in low EDTA-TE Buffer
Vibrio vulnificus	NZRM 2506^T	20 ng/µL		None	diluted in low EDTA-TE Buffer
Yersinia enterocolitica	NZRM 767	20 ng/µL		None	diluted in low EDTA-TE Buffer
Yersinia pseudotuberculosis	NZRM 768^T	20 ng/µL		None	diluted in low EDTA-TE Buffer

[a] Probes in italics were visible but not assigned by the band matching. Probes in bold were not expected based on the taxa of the DNA extracts

Reference List

Aabenhus, R., S. L. W. On, B. L. Siemer, H. Permin, and L. P. Andersen. 2005. "Delineation of *Campylobacter concisus* genomospecies by amplified fragment length polymorphism analysis and correlation of results with clinical data." *J Clin Microbiol* 43 (10):5091-6. doi: 10.1128/jcm.43.10.5091-5096.2005.

Abdelbaqi, K., A. Buissonniere, V. Prouzet-Mauleon, J. Gresser, I. Wesley, F. Megraud, and A. Menard. 2007. "Development of a real-time fluorescence resonance energy transfer PCR to detect *Arcobacter* species." *J Clin Microbiol* 45 (9):3015-21. doi: 10.1128/jcm.00256-07.

Adlam, S. B., S. Perera, R. J. Lake, D. M. Campbell, J. A. Williman, and M. G. Baker. 2011. "Acute gastrointestinal illness in New Zealand: a community study." *Epidemiol Infect* 139 (2):302-8. doi: 10.1017/s0950268810000932.

Agren, J., A. Sundstrom, T. Hafstrom, and B. Segerman. 2012. "Gegenees: fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups." *PLoS One* 7 (6):e39107. doi: 10.1371/journal.pone.0039107.

Aken, B. L., S. Ayling, D. Barrell, L. Clarke, V. Curwen, S. Fairley, J. Fernandez Banet, K. Billis, C. Garcia Giron, T. Hourlier, K. Howe, A. Kahari, F. Kokocinski, F. J. Martin, D. N. Murphy, R. Nag, M. Ruffier, M. Schuster, Y. A. Tang, J. H. Vogel, S. White, A. Zadissa, P. Flicek, and S. M. J. Searle. 2016. "The Ensembl gene annotation system." *Database (Oxford)* 2016. doi: 10.1093/database/baw093.

Al-Soud, W. A., and P. Radstrom. 2001. "Purification and characterization of PCR-inhibitory components in blood cells." *J Clin Microbiol* 39 (2):485-93. doi: 10.1128/jcm.39.2.485-493.2001.

Alain, K., N. Callac, M. Guegan, F. Lesongeur, P. Crassous, M. A. Cambon-Bonavita, J. Querellou, and D. Prieur. 2009. "*Nautilia abyssi* sp. nov., a thermophilic, chemolithoautotrophic, sulfur-reducing bacterium isolated from an East Pacific Rise hydrothermal vent." *Int J Syst Evol Microbiol* 59 (Pt 6):1310-5. doi: 10.1099/ijs.0.005454-0.

Alain, K., J. Querellou, F. Lesongeur, P. Pignet, P. Crassous, G. Raguenes, V. Cueff, and M. A. Cambon-Bonavita. 2002. "*Caminibacter hydrogeniphilus* gen. nov., sp. nov., a novel thermophilic, hydrogen-oxidizing bacterium isolated from an East Pacific Rise hydrothermal vent." *Int J Syst Evol Microbiol* 52 (Pt 4):1317-23. doi: 10.1099/ijs.0.02195-0.

Alispahic, M., K. Hummel, D. Jandreski-Cvetkovic, K. Nobauer, E. Razzazi-Fazeli, M. Hess, and C. Hess. 2010. "Species-specific identification and differentiation of *Arcobacter*, *Helicobacter* and *Campylobacter* by full-spectral matrix-associated laser

desorption/ionization time of flight mass spectrometry analysis." *J Med Microbiol* 59 (Pt 3):295-301. doi: 10.1099/jmm.0.016576-0.

Altermann, E., J. Lu, and A. McCulloch. 2017. "GAMOLA2, a comprehensive software package for the annotation and curation of draft and complete microbial genomes." *Front Microbiol* 8:346. doi: 10.3389/fmicb.2017.00346.

Amar, C. F. L., C. L. East, J. Gray, M. Iturriza-Gomara, E. A. Maclure, and J. McLauchlin. 2007. "Detection by PCR of eight groups of enteric pathogens in 4,627 faecal samples: re-examination of the English case-control Infectious Intestinal Disease Study (1993-1996)." *Eur J Clin Microbiol Infect Dis* 26 (5):311-23. doi: 10.1007/s10096-007-0290-8.

Anderson, N. W., B. W. Buchan, and N. A. Ledeboer. 2014. "Comparison of the BD MAX enteric bacterial panel to routine culture methods for detection of *Campylobacter*, enterohemorrhagic *Escherichia coli* (O157), *Salmonella*, and *Shigella* isolates in preserved stool specimens." *J Clin Microbiol* 52 (4):1222-4. doi: 10.1128/jcm.03099-13.

Anonymous. 1949. gastroenteritis. In *Blakiston's New Gould Medical Dictionary*, edited by H. W. Jones, N. L. Hoerr and A. Osol. Philadelphia, USA: The Blakiston Company.

Anonymous. 1994. "Infection with *Helicobacter pylori*." *IARC Monogr Eval Carcinog Risks Hum* 61:177-240.

Anonymous. 2012. The Global View of Campylobacteriosis. Report of an Expert Consultation Utrecht, Netherlands, 9-11 July 2012. Geneva, Switzerland: World Health Organisation.

Arrieta, M. C., L. T. Stiemsma, N. Amenyogbe, E. M. Brown, and B. Finlay. 2014. "The intestinal microbiome in early life: health and disease." *Front Immunol* 5:427. doi: 10.3389/fimmu.2014.00427.

Auch, A. F., H. P. Klenk, and M. Goker. 2010. "Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs." *Stand Genomic Sci* 2 (1):142-8. doi: 10.4056/sigs.541628.

Auch, A. F., M. von Jan, H. P. Klenk, and M. Goker. 2010. "Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison." *Stand Genomic Sci* 2 (1):117-34. doi: 10.4056/sigs.531120.

Aziz, R. K., D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, F. Meyer, G. J. Olsen, R. Olson, A. L. Osterman, R. A. Overbeek, L. K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke, and O. Zagnitko. 2008. "The RAST Server: rapid annotations using subsystems technology." *BMC Genomics* 9:75. doi: 10.1186/1471-2164-9-75.

Badger, J. H., and G. J. Olsen. 1999. "CRITICA: coding region identification tool invoking comparative analysis." *Mol Biol Evol* 16 (4):512-24. doi: 10.1093/oxfordjournals.molbev.a026133.

Baele, M., A. Decostere, P. Vandamme, K. Van den Bulck, I. Gruntar, J. Mehle, J. Mast, R. Ducatelle, and F. Haesebrouck. 2008. "*Helicobacter baculiformis* sp. nov., isolated from

feline stomach mucosa." *Int J Syst Evol Microbiol* 58 (Pt 2):357-64. doi: 10.1099/ijs.0.65152-0.

Baig, A., A. McNally, S. Dunn, K. H. Paszkiewicz, J. Corander, and G. Manning. 2015. "Genetic import and phenotype specific alleles associated with hyper-invasion in *Campylobacter jejuni*." *BMC Genomics* 16:852. doi: 10.1186/s12864-015-2087-y.

Barrett, J., and M. Brown. 2016. "Travellers' diarrhoea." *BMJ* 353:i1937. doi: 10.1136/bmj.i1937.

Bastyns, K., S. Chapelle, P. Vandamme, H. Goossens, and R. De Wachter. 1995. "Specific detection of *Campylobacter concisus* by PCR amplification of 23S rDNA areas." *Mol Cell Probes* 9 (4):247-50.

Behjati, S., and P. S. Tarpey. 2013. "What is next generation sequencing?" *Arch Dis Child Educ Pract Ed* 98 (6):236-8. doi: 10.1136/archdischild-2013-304340.

Ben Zakour, N. L., S. A. Beatson, A. H. van den Broek, K. L. Thoday, and J. R. Fitzgerald. 2012. "Comparative genomics of the *Staphylococcus intermedius* group of animal pathogens." *Front Cell Infect Microbiol* 2:44. doi: 10.3389/fcimb.2012.00044.

Benard-Slagter, A., I. Zondervan, K. de Groot, F. Ghazavi, V. Sarhadi, P. Van Vlierberghe, B. De Moerloose, C. Schwab, K. Vettenranta, C. J. Harrison, S. Knuutila, J. Schouten, T. Lammens, and S. Savola. 2017. "Digital multiplex ligation-dependent probe amplification for detection of key copy number alterations in T- and B-cell lymphoblastic leukemia." *J Mol Diagn* 19 (5):659-672. doi: 10.1016/j.jmoldx.2017.05.004.

Benedict, M. N., J. R. Henriksen, W. W. Metcalf, R. J. Whitaker, and N. D. Price. 2014. "ITEP: an integrated toolkit for exploration of microbial pan-genomes." *BMC Genomics* 15:8. doi: 10.1186/1471-2164-15-8.

Bennett, J. S., K. A. Jolley, S. G. Earle, C. Corton, S. D. Bentley, J. Parkhill, and M. C. J. Maiden. 2012. "A genomic approach to bacterial taxonomy: an examination and proposed reclassification of species within the genus *Neisseria*." *Microbiology* 158 (Pt 6):1570-80. doi: 10.1099/mic.0.056077-0.

Bennett, J. S., K. A. Jolley, and M. C. J. Maiden. 2013. "Genome sequence analyses show that *Neisseria oralis* is the same species as *'Neisseria mucosa* var. *heidelbergensis*'." *Int J Syst Evol Microbiol* 63 (Pt 10):3920-6. doi: 10.1099/ijs.0.052431-0.

Benson, D. A., M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. 2017. "GenBank." *Nucleic Acids Res* 45 (D1):D37-D42. doi: 10.1093/nar/gkw1070.

Bergval, I., S. Sengstake, N. Brankova, V. Levterova, E. Abadia, N. Tadumaze, N. Bablishvili, M. Akhalaia, K. Tuin, A. Schuitema, S. Panaiotov, E. Bachiyska, T. Kantardjiev, R. de Zwaan, A. Schurch, D. van Soolingen, A. van 't Hoog, F. Cobelens, R. Aspindzelashvili, C. Sola, P. Klatser, and R. Anthony. 2012. "Combined species identification, genotyping, and drug resistance detection of *Mycobacterium tuberculosis* cultures by MLPA on a bead-based array." *PLoS One* 7 (8):e43240. doi: 10.1371/journal.pone.0043240.

Bessede, E., O. Solecki, E. Sifre, L. Labadi, and F. Megraud. 2011. "Identification of *Campylobacter* species and related organisms by matrix assisted laser desorption ionization-time of flight (MALDI-TOF) mass spectrometry." *Clin Microbiol Infect* 17 (11):1735-9. doi: 10.1111/j.1469-0691.2011.03468.x.

Biomatters Ltd. 2015. "Geneious 8.1 Manual." In: Biomatters Ltd (accessed 21 April 2015).

Blackett, K. L., S. S. Siddhi, S. Cleary, H. Steed, M. H. Miller, S. Macfarlane, G. T. Macfarlane, and J. F. Dillon. 2013. "Oesophageal bacterial biofilm changes in gastro-oesophageal reflux disease, Barrett's and oesophageal carcinoma: association or causality?" *Aliment Pharmacol Ther* 37 (11):1084-92. doi: 10.1111/apt.12317.

Blacklow, N. R., and H. B. Greenberg. 1991. "Viral gastroenteritis." *N Engl J Med* 325 (4):252-64. doi: 10.1056/nejm199107253250406.

Blom, J., S. P. Albaum, D. Doppmeier, A. Puhler, F. J. Vorholter, M. Zakrzewski, and A. Goesmann. 2009. "EDGAR: a software framework for the comparative analysis of prokaryotic genomes." *BMC Bioinformatics* 10:154. doi: 10.1186/1471-2105-10-154.

Boga, J. A., S. Melon, I. Nicieza, I. de Diego, M. Villar, F. Parra, and M. de Ona. 2004. "Etiology of sporadic cases of pediatric acute gastroenteritis in Asturias, Spain, and genotyping and characterization of norovirus strains involved." *J Clin Microbiol* 42 (6):2668-74. doi: 10.1128/jcm.42.6.2668-2674.2004.

Bojanic, K. 2017. "Source of Kruno's isolates/genomes." email 20/3/17.

Bojanic, K., A. C. Midwinter, J. C. Marshall, L. E. Rogers, P. J. Biggs, and E. Acke. 2017. "Isolation of *Campylobacter* spp. from client-owned dogs and cats, and retail raw meat pet food in the Manawatu, New Zealand." *Zoonoses Public Health* 64 (6):438-449. doi: 10.1111/zph.12323.

Bresee, J. S., R. Marcus, R. A. Venezia, W. E. Keene, D. Morse, M. Thanassi, P. Brunett, S. Bulens, R. S. Beard, L. A. Dauphin, L. Slutsker, C. Bopp, M. Eberhard, A. Hall, J. Vinje, S. S. Monroe, R. I. Glass, and US Acute Gastroenteritis Etiology (AGE) Study Team. 2012. "The etiology of severe acute gastroenteritis among adults visiting emergency departments in the United States." *J Infect Dis* 205 (9):1374-81. doi: 10.1093/infdis/jis206.

Brittnacher, M. J., C. Fong, H. S. Hayden, M. A. Jacobs, M. Radey, and L. Rohmer. 2011. "PGAT: a multistrain analysis resource for microbial genomes." *Bioinformatics* 27 (17):2429-30. doi: 10.1093/bioinformatics/btr418.

Broczyk, A., S. Thompson, D. Smith, and H. Lior. 1987. "Water-borne outbreak of *Campylobacter laridis*-associated gastroenteritis." *Lancet* 1 (8525):164-5. doi: 10.1016/S0140-6736(87)92003-4.

Brondz, I., and I. Olsen. 1991. "Multivariate analyses of cellular fatty acids in *Bacteroides*, *Prevotella*, *Porphyromonas*, *Wolinella*, and *Campylobacter* spp." *J Clin Microbiol* 29 (1):183-9.

Bruno, J. G., T. Phillips, M. P. Carrillo, and R. Crowell. 2009. "Plastic-adherent DNA aptamer-magnetic bead and quantum dot sandwich assay for *Campylobacter* detection." *J Fluoresc* 19 (3):427-35. doi: 10.1007/s10895-008-0429-8.

Brynildsrud, O., J. Bohlin, L. Scheffer, and V. Eldholm. 2016. "Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary." *Genome Biol* 17 (1):238. doi: 10.1186/s13059-016-1108-8.

Bull, M. J., J. R. Marchesi, P. Vandamme, S. Plummer, and E. Mahenthiralingam. 2012. "Minimum taxonomic criteria for bacterial genome sequence depositions and announcements." *J Microbiol Methods* 89 (1):18-21. doi: 10.1016/j.mimet.2012.02.008.

Bullman, S., D. Corcoran, J. O'Leary, D. Byrne, B. Lucey, and R. D. Sleator. 2012. "Epsilonproteobacteria in humans, New Zealand." *Emerg Infect Dis* 18 (10):1709-10; author reply 1710-1. doi: 10.3201/eid1810.120369.

Bullman, S., D. Corcoran, J. O'Leary, B. Lucey, D. Byrne, and R. D. Sleator. 2011. "*Campylobacter ureolyticus*: an emerging gastrointestinal pathogen?" *FEMS Immunol Med Microbiol* 61 (2):228-30. doi: 10.1111/j.1574-695X.2010.00760.x.

Bullman, S., A. Lucid, D. Corcoran, R. D. Sleator, and B. Lucey. 2013. "Genomic investigation into strain heterogeneity and pathogenic potential of the emerging gastrointestinal pathogen *Campylobacter ureolyticus*." *PLoS One* 8 (8):e71515. doi: 10.1371/journal.pone.0071515.

Burnens, A. P., J. Stanley, R. Morgenstern, and J. Nicolet. 1994. "Gastroenteritis associated with *Helicobacter pullorum*." *Lancet* 344 (8936):1569-70. doi: 10.1016/S0140-6736(94)90376-X.

Caceres, A., I. Munoz, G. Iraola, F. Diaz-Viraque, and L. Collado. 2017. "*Campylobacter ornithocola* sp. nov., a novel member of the *Campylobacter lari* group isolated from wild bird faecal samples." *Int J Syst Evol Microbiol* 67:1643-49. doi: 10.1099/ijsem.0.001822.

Canadian Agency for Drugs and Technologies in Health. 2015. Stool Antigen Tests for *Helicobacter pylori* Infection: A Review of Clinical and Cost-Effectiveness and Guidelines. In *CADTH Rapid Response Reports*. Ottawa (ON): Canadian Agency for Drugs and Technologies in Health

Carter, M. Q. 2017. "Decoding the ecological function of accessory genome." *Trends Microbiol* 25 (1):6-8. doi: 10.1016/j.tim.2016.11.012.

Ceelen, L., A. Decostere, G. Verschraegen, R. Ducatelle, and F. Haesebrouck. 2005. "Prevalence of *Helicobacter pullorum* among patients with gastrointestinal disease and clinically healthy persons." *J Clin Microbiol* 43 (6):2984-6. doi: 10.1128/jcm.43.6.2984-2986.2005.

Chaban, B., K. M. Musil, C. G. Himsworth, and J. E. Hill. 2009. "Development of *cpn60*-based real-time quantitative PCR assays for the detection of 14 *Campylobacter* species and application to screening of canine fecal samples." *Appl Environ Microbiol* 75 (10):3055-61. doi: 10.1128/aem.00101-09.

Chaban, B., M. Ngeleka, and J. E. Hill. 2010. "Detection and quantification of 14 *Campylobacter* species in pet dogs reveals an increase in species richness in feces of diarrheic animals." *BMC Microbiol* 10:73. doi: 10.1186/1471-2180-10-73.

Chaudhari, N. M., V. K. Gupta, and C. Dutta. 2016. "BPGA- an ultra-fast pan-genome analysis pipeline." *Sci Rep* 6:24373. doi: 10.1038/srep24373.

Chen, I. M. A., V. M. Markowitz, K. Palaniappan, E. Szeto, K. Chu, J. Huang, A. Ratner, M. Pillay, M. Hadjithomas, M. Huntemann, N. Mikhailova, G. Ovchinnikova, N. N. Ivanova, and N. C. Kyrpides. 2016. "Supporting community annotation and user collaboration in the integrated microbial genomes (IMG) system." *BMC Genomics* 17:307. doi: 10.1186/s12864-016-2629-y.

Chen, P. E., and B. J. Shapiro. 2015. "The advent of genome-wide association studies for bacteria." *Curr Opin Microbiol* 25:17-24. doi: 10.1016/j.mib.2015.03.002.

Cheun, H. I., S. H. Cho, J. H. Lee, Y. Y. Lim, J. H. Jeon, J. R. Yu, T. S. Kim, W. J. Lee, S. H. Cho, D. Y. Lee, M. S. Park, H. S. Jeong, D. S. Chen, Y. M. Ji, and M. H. Kwon. 2010. "Infection status of hospitalized diarrheal patients with gastrointestinal protozoa, bacteria, and viruses in the Republic of Korea." *Korean J Parasitol* 48 (2):113-20. doi: 10.3347/kjp.2010.48.2.113.

Chun, J., and F. A. Rainey. 2014. "Integrating genomics into the taxonomy and systematics of the *Bacteria* and *Archaea*." *Int J Syst Evol Microbiol* 64 (Pt 2):316-24. doi: 10.1099/ijs.0.054171-0.

Chung, B., G. W. Shin, J. Na, M. H. Oh, and G. Y. Jung. 2012. "Multiplex quantitative foodborne pathogen detection using high resolution CE-SSCP coupled stuffer-free multiplex ligation-dependent probe amplification." *Electrophoresis* 33 (9-10):1477-81. doi: 10.1002/elps.201100615.

Chung, H. K. L., A. Tay, S. Octavia, J. Chen, F. Liu, R. Ma, R. Lan, S. M. Riordan, M. C. Grimm, and L. Zhang. 2016. "Genome analysis of *Campylobacter concisus* strains from patients with inflammatory bowel disease and gastroenteritis provides new insights into pathogenicity." *Sci Rep* 6:38442. doi: 10.1038/srep38442.

Cody, A. J., J. E. Bray, K. A. Jolley, N. D. McCarthy, and M. C. J. Maiden. 2017. "Core genome multilocus sequence typing scheme for stable, comparative analyses of *Campylobacter jejuni* and *C. coli* human disease isolates." *J Clin Microbiol* 55 (7):2086-2097. doi: 10.1128/jcm.00080-17.

Collado, L., and M. J. Figueras. 2011. "Taxonomy, epidemiology, and clinical relevance of the genus *Arcobacter*." *Clin Microbiol Rev* 24 (1):174-92. doi: 10.1128/cmr.00034-10.

Collado, L., M. Gutierrez, M. Gonzalez, and H. Fernandez. 2013. "Assessment of the prevalence and diversity of emergent campylobacteria in human stool samples using a combination of traditional and molecular methods." *Diagn Microbiol Infect Dis* 75 (4):434-6. doi: 10.1016/j.diagmicrobio.2012.12.006.

Collins, F. S., M. Morgan, and A. Patrinos. 2003. "The Human Genome Project: lessons from large-scale biology." *Science* 300 (5617):286-90. doi: 10.1126/science.1084564.

Colomba, C., S. De Grazia, G. M. Giammanco, L. Saporito, F. Scarlata, L. Titone, and S. Arista. 2006. "Viral gastroenteritis in children hospitalised in Sicily, Italy." *Eur J Clin Microbiol Infect Dis* 25 (9):570-5. doi: 10.1007/s10096-006-0188-x.

Cone, R. W., A. C. Hobson, and M. L. W. Huang. 1992. "Coamplified positive control detects inhibition of polymerase chain reactions." *J Clin Microbiol* 30 (12):3185-9.

Contreras-Moreira, B., and P. Vinuesa. 2013. "GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis." *Appl Environ Microbiol* 79 (24):7696-701. doi: 10.1128/aem.02411-13.

Cornelius, A. J., S. Chambers, J. Aitken, S. M. Brandt, B. Horn, and S. L. W. On. 2012. "Epsilonproteobacteria in humans, New Zealand." *Emerg Infect Dis* 18 (3):510-2. doi: 10.3201/eid1803.110875.

Cornelius, A. J., W. G. Miller, A. J. Lastovica, S. L. W. On, N. P. French, O. Vandenberg, and P. J. Biggs. 2017. "Complete genome sequence of *Campylobacter concisus* ATCC 33237^{T} and draft genome sequences for an additional eight well-characterized *C. concisus* strains." *Genome Announc* 5 (29):e00711-17. doi: 10.1128/genomeA.00711-17.

Cornelius, A. J., O. Vandenberg, B. Robson, B. J. Gilpin, S. M. Brandt, P. Scholes, D. Martiny, P. E. Carter, P. van Vught, J. Schouten, and S. L. W. On. 2014. "Same-day subtyping of *Campylobacter jejuni* and *C. coli* isolates by use of multiplex ligation-dependent probe amplification-binary typing." *J Clin Microbiol* 52 (9):3345-50. doi: 10.1128/jcm.00815-14.

Coupland, L. J., I. McElarney, E. Meader, K. Cowley, L. Alcock, J. Naunton, and J. Gray. 2013. "Simultaneous detection of viral and bacterial enteric pathogens using the Seeplex(R) Diarrhea ACE detection system." *Epidemiol Infect* 141 (10):2111-21. doi: 10.1017/s0950268812002622.

Cowan, S. T. 1965. "Principles and practice of bacterial taxonomy - a forward look." *J Gen Microbiol* 39:143-53. doi: 10.1099/00221287-39-1-143.

Cox, M. P., D. A. Peterson, and P. J. Biggs. 2010. "SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data." *BMC Bioinformatics* 11:485. doi: 10.1186/1471-2105-11-485.

Dark, M. J. 2013. "Whole-genome sequencing in bacteriology: state of the art." *Infect Drug Resist* 6:115-23. doi: 10.2147/idr.s35710.

de Boer, R. F., A. Ott, B. Kesztyus, and A. M. D. Kooistra-Smid. 2010. "Improved detection of five major gastrointestinal pathogens by use of a molecular screening approach." *J Clin Microbiol* 48 (11):4140-6. doi: 10.1128/jcm.01124-10.

De Smet, S., P. Vandamme, L. De Zutter, S. L. W. On, L. Douidah, and K. Houf. 2011. "*Arcobacter trophiarum* sp. nov., isolated from fattening pigs." *Int J Syst Evol Microbiol* 61 (Pt 2):356-61. doi: 10.1099/ijs.0.022665-0.

de Vries, J. J., N. L. A. Arents, and W. L. Manson. 2008. "*Campylobacter* species isolated from extra-oro-intestinal abscesses: a report of four cases and literature review." *Eur J Clin Microbiol Infect Dis* 27 (11):1119-23. doi: 10.1007/s10096-008-0550-2.

de Wit, M. A. S., M. P. G. Koopmans, L. M. Kortbeek, N. J. van Leeuwen, A. I. M. Bartelds, and Y. T. H. P. van Duynhoven. 2001. "Gastroenteritis in sentinel general practices, the Netherlands." *Emerg Infect Dis* 7 (1):82-91. doi: 10.3201/eid0701.700082.

de Wit, M. A. S., M. P. G. Koopmans, L. M. Kortbeek, N. J. van Leeuwen, J. Vinje, and Y. T. H. P. van Duynhoven. 2001. "Etiology of gastroenteritis in sentinel general practices in the Netherlands." *Clin Infect Dis* 33 (3):280-288. doi: 10.1086/321875.

de Wit, M. A. S., M. P. G. Koopmans, L. M. Kortbeek, W. J. B. Wannet, J. Vinje, F. van Leusden, A. I. M. Bartelds, and Y. T. H. P. van duynhoven. 2001. "Sensor, a population-based cohort study on gastroenteritis in the Netherlands: incidence and etiology." *Am J Epidemiol* 154 (7):666-74. doi: 10.1093/aje/154.7.666.

Debruyne, L., T. Broman, S. Bergstrom, B. Olsen, S. L. W. On, and P. Vandamme. 2010a. "*Campylobacter subantarcticus* sp. nov., isolated from birds in the sub-Antarctic region." *Int J Syst Evol Microbiol* 60 (Pt 4):815-9. doi: 10.1099/ijs.0.011056-0.

Debruyne, L., T. Broman, S. Bergstrom, B. Olsen, S. L. W. On, and P. Vandamme. 2010b. "*Campylobacter volucris* sp. nov., isolated from black-headed gulls (*Larus ridibundus*)." *Int J Syst Evol Microbiol* 60 (Pt 8):1870-5. doi: 10.1099/ijs.0.013748-0.

Debruyne, L., D. Gevers, and P. Vandamme. 2008. "Taxonomy of the family *Campylobacteraceae*." In *Campylobacter*, edited by I. Nachamkin, C. M. Szymanski and M. J. Blaser, 3-25. Washington DC: ASM Press.

Debruyne, L., S. L. W. On, E. De Brandt, and P. Vandamme. 2009. "Novel *Campylobacter lari*-like bacteria from humans and molluscs: description of *Campylobacter peloridis* sp. nov., *Campylobacter lari* subsp. *concheus* subsp. nov. and *Campylobacter lari* subsp. *lari* subsp. nov." *Int J Syst Evol Microbiol* 59 (Pt 5):1126-32. doi: 10.1099/ijs.0.000851-0.

Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg. 1999. "Improved microbial gene identification with GLIMMER." *Nucleic Acids Res* 27 (23):4636-41. doi: 10.1093/nar/27.23.4636.

Deshpande, N. P., N. O. Kaakoush, H. Mitchell, K. Janitz, M. J. Raftery, S. S. Li, and M. R. Wilkins. 2011. "Sequencing and validation of the genome of a *Campylobacter concisus* reveals intra-species diversity." *PLoS One* 6 (7):e22170. doi: 10.1371/journal.pone.0022170.

Deshpande, N. P., N. O. Kaakoush, M. R. Wilkins, and H. M. Mitchell. 2013. "Comparative genomics of *Campylobacter concisus* isolates reveals genetic diversity and provides insights into disease association." *BMC Genomics* 14:585. doi: 10.1186/1471-2164-14-585.

Dewhirst, F. E., J. G. Fox, and S. L. W. On. 2000. "Recommended minimal standards for describing new species of the genus *Helicobacter*." *Int J Syst Evol Microbiol* 50 (Pt 6):2231-7. doi: 10.1099/00207713-50-6-2231.

Dewhirst, F. E., Z. Shen, M. S. Scimeca, L. N. Stokes, T. Boumenna, T. Chen, B. J. Paster, and J. G. Fox. 2005. "Discordant 16S and 23S rRNA gene phylogenies for the genus

Helicobacter: implications for phylogenetic inference and systematics." *J Bacteriol* 187 (17):6106-18. doi: 10.1128/jb.187.17.6106-6118.2005.

Ding, W., F. Baumdicker, and R. A. Neher. 2018. "panX: pan-genome analysis and exploration." *Nucleic Acids Res* 46 (1):e5. doi: 10.1093/nar/gkx977.

Donachie, S. P., J. P. Bowman, S. L. W. On, and M. Alam. 2005. "*Arcobacter halophilus* sp. nov., the first obligate halophile in the genus *Arcobacter*." *Int J Syst Evol Microbiol* 55 (Pt 3):1271-7. doi: 10.1099/ijs.0.63581-0.

Donatin, E., S. Buffet, Q. Leroy, D. Raoult, and M. Drancourt. 2013. "A DNA microarray for the versatile diagnosis of infectious diarrhea." *APMIS* 121 (7):634-42. doi: 10.1111/apm.12081.

Dunbar, S. A., H. Zhang, and Y. W. Tang. 2013. "Advanced techniques for detection and identification of microbial agents of gastroenteritis." *Clin Lab Med* 33 (3):527-52. doi: 10.1016/j.cll.2013.03.003.

El-Metwally, S., T. Hamza, M. Zakaria, and M. Helmy. 2013. "Next-generation sequence assembly: four stages of data processing and computational challenges." *PLoS Comput Biol* 9 (12):e1003345. doi: 10.1371/journal.pcbi.1003345.

Elliott, E. J. 2007. "Acute gastroenteritis in children." *BMJ* 334 (7583):35-40. doi: 10.1136/bmj.39036.406169.80.

Engberg, J., D. D. Bang, R. Aabenhus, F. M. Aarestrup, V. Fussing, and P. Gerner-Smidt. 2005. "*Campylobacter concisus*: an evaluation of certain phenotypic and genotypic characteristics." *Clin Microbiol Infect* 11 (4):288-95. doi: 10.1111/j.1469-0691.2005.01111.x.

Engberg, J., S. L. W. On, C. S. Harrington, and P. Gerner-Smidt. 2000. "Prevalence of *Campylobacter*, *Arcobacter*, *Helicobacter*, and *Sutterella* spp. in human fecal samples as estimated by a reevaluation of isolation methods for campylobacters." *J Clin Microbiol* 38 (1):286-91.

Etoh, Y., F. E. Dewhirst, B. J. Paster, A. Yamamoto, and N. Goto. 1993. "*Campylobacter showae* sp. nov., isolated from the human oral cavity." *Int J Syst Bacteriol* 43 (4):631-9. doi: 10.1099/00207713-43-4-631.

Etoh, Y., A. Yamamoto, and N. Goto. 1998. "Intervening sequences in 16S rRNA genes of *Campylobacter* sp.: diversity of nucleotide sequences and uniformity of location." *Microbiol Immunol* 42 (3):241-3. doi: 10.1111/j.1348-0421.1998.tb02278.x.

Euzéby, J. P. 1997. "List of bacterial names with standing in nomenclature: a folder available on the internet." *Int J Syst Bacteriol* 47 (2):590-592. doi: 10.1099/00207713-47-2-590.

Evans, A. S. 1976. "Causation and disease: the Henle-Koch postulates revisited." *Yale J Biol Med* 49 (2):175-95.

Facey, P. D., G. Meric, M. D. Hitchings, J. A. Pachebat, M. J. Hegarty, X. Chen, L. V. A. Morgan, J. E. Hoeppner, M. M. A. Whitten, W. D. J. Kirk, P. J. Dyson, S. K. Sheppard, and R. Del Sol. 2015. "Draft genomes, phylogenetic reconstruction, and comparative

genomics of two novel cohabiting bacterial symbionts isolated from *Frankliniella occidentalis*." *Genome Biol Evol* 7 (8):2188-202. doi: 10.1093/gbe/evv136.

Fachmann, M. S. R., M. H. Josefsen, J. Hoorfar, M. T. Nielsen, and C. Lofstrom. 2015. "Cost-effective optimization of real-time PCR-based detection of *Campylobacter* and *Salmonella* with inhibitor tolerant DNA polymerases." *J Appl Microbiol* 119 (5):1391-402. doi: 10.1111/jam.12937.

Ferreira, S., C. Julio, J. A. Queiroz, F. C. Domingues, and M. Oleastro. 2014. "Molecular diagnosis of *Arcobacter* and *Campylobacter* in diarrhoeal samples among Portuguese patients." *Diagn Microbiol Infect Dis* 78 (3):220-5. doi: 10.1016/j.diagmicrobio.2013.11.021.

Fiedoruk, K., T. Daniluk, D. Rozkiewicz, M. L. Zaremba, E. Oldak, M. Sciepuk, and K. Leszczynska. 2015. "Conventional and molecular methods in the diagnosis of community-acquired diarrhoea in children under 5 years of age from the north-eastern region of Poland." *Int J Infect Dis* 37:145-51. doi: 10.1016/j.ijid.2015.06.028.

Figura, N., P. Guglielmetti, A. Zanchi, N. Partini, D. Armellini, P. F. Bayeli, M. Bugnoli, and S. Verdiani. 1993. "Two cases of *Campylobacter mucosalis* enteritis in children." *J Clin Microbiol* 31 (3):727-8.

Finn, R. D., T. K. Attwood, P. C. Babbitt, A. Bateman, P. Bork, A. J. Bridge, H. Y. Chang, Z. Dosztanyi, S. El-Gebali, M. Fraser, J. Gough, D. Haft, G. L. Holliday, H. Huang, X. Huang, I. Letunic, R. Lopez, S. Lu, A. Marchler-Bauer, H. Mi, J. Mistry, D. A. Natale, M. Necci, G. Nuka, C. A. Orengo, Y. Park, S. Pesseat, D. Piovesan, S. C. Potter, N. D. Rawlings, N. Redaschi, L. Richardson, C. Rivoire, A. Sangrador-Vegas, C. Sigrist, I. Sillitoe, B. Smithers, S. Squizzato, G. Sutton, N. Thanki, P. D. Thomas, S. C. E. Tosatto, C. H. Wu, I. Xenarios, L. S. Yeh, S. Y. Young, and A. L. Mitchell. 2017. "InterPro in 2017-beyond protein family and domain annotations." *Nucleic Acids Res* 45 (D1):D190-d199. doi: 10.1093/nar/gkw1107.

Finn, R. D., A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, and M. Punta. 2014. "Pfam: the protein families database." *Nucleic Acids Res* 42 (Database issue):D222-30. doi: 10.1093/nar/gkt1223.

Finn, R. D., J. Clements, and S. R. Eddy. 2011. "HMMER web server: interactive sequence similarity searching." *Nucleic Acids Res* 39 (Web Server issue):W29-37. doi: 10.1093/nar/gkr367.

Finn, R. D., P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, and A. Bateman. 2016. "The Pfam protein families database: towards a more sustainable future." *Nucleic Acids Res* 44 (D1):D279-85. doi: 10.1093/nar/gkv1344.

Firth, C., and W. I. Lipkin. 2013. "The genomics of emerging pathogens." *Annu Rev Genomics Hum Genet* 14:281-300. doi: 10.1146/annurev-genom-091212-153446.

Fitzgerald, C., M. Patrick, A. Gonzalez, J. Akin, C. R. Polage, K. Wymore, L. Gillim-Ross, K. Xavier, J. Sadlowski, J. Monahan, S. Hurd, S. Dahlberg, R. Jerris, R. Watson, M. Santovenia, D. Mitchell, C. Harrison, M. Tobin-D'Angelo, M. DeMartino, M. Pentella, J. Razeq, C. Leonard, C. Jung, R. Achong-Bowe, Y. Evans, D. Jain, B. Juni, F. Leano, T. Robinson, K. Smith, R. M. Gittelman, C. Garrigan, I. Nachamkin, and Campylobacter Diagnostics Study working Group. 2016. "Multicenter evaluation of clinical diagnostic methods for detection and isolation of *Campylobacter* spp. from stool." *J Clin Microbiol* 54 (5):1209-15. doi: 10.1128/jcm.01925-15.

Foodborne Disease Burden Epidemiology Reference Group 2007-2015. 2015. WHO estimates of the global burden of foodborne deseases: Foodborne Disease Burden Epidemiology Reference Group 2007-2015. edited by World Health Organisation. Geneva, Switzerland: World Health Organisation.

Forde, B. M., and P. W. O'Toole. 2013. "Next-generation sequencing technologies and their impact on microbial genomics." *Brief Funct Genomics* 12 (5):440-53. doi: 10.1093/bfgp/els062.

Forouzan, E., M. S. M. Maleki, A. A. Karkhane, and B. Yakhchali. 2017. "Evaluation of nine popular *de novo* assemblers in microbial genome assembly." *J Microbiol Methods* 143:32-37. doi: 10.1016/j.mimet.2017.09.008.

Foster, G., B. Holmes, A. G. Steigerwalt, P. A. Lawson, P. Thorne, D. E. Byrer, H. M. Ross, J. Xerry, P. M. Thompson, and M. D. Collins. 2004. "*Campylobacter insulaenigrae* sp. nov., isolated from marine mammals." *Int J Syst Evol Microbiol* 54 (Pt 6):2369-73. doi: 10.1099/ijs.0.63147-0.

Fowsantear, W., E. Argo, C. Pattinson, and P. Cash. 2014. "Comparative proteomics of *Helicobacter* species: the discrimination of gastric and enterohepatic *Helicobacter* species." *J Proteomics* 97:245-55. doi: 10.1016/j.jprot.2013.07.016.

Fox, J. G., Terry Chilvers, C. Stewart Goodwin, N. S. Taylor, Paul Edmonds, Lindsay I. Sly, and Don J. Brenner. 1989. "*Campylobacter mustelae*, a new species resulting from the elevation of *Campylobacter pylori* subsp. *mustelae* to species status." *Int J Syst Bacteriol* 39 (3):301-303. doi: doi:10.1099/00207713-39-3-301.

Friesema, I. H. M., R. F. de Boer, E. Duizer, L. M. Kortbeek, D. W. Notermans, O. F. Norbruis, D. D. L. Bezemer, H. van Heerbeek, R. N. J. van Andel, J. G. van Enk, P. L. A. Fraaij, M. P. G. Koopmans, A. M. D. Kooistra-Smid, and Y. T. H. P. van duynhoven. 2012. "Etiology of acute gastroenteritis in children requiring hospitalization in the Netherlands." *Eur J Clin Microbiol Infect Dis* 31 (4):405-15. doi: 10.1007/s10096-011-1320-0.

Frishman, D., A. Mironov, H. W. Mewes, and M. Gelfand. 1998. "Combining diverse evidence for gene recognition in completely sequenced bacterial genomes." *Nucleic Acids Res* 26 (12):2941-7. doi: 10.1093/nar/26.12.2941.

Gadiel, D., and P. Abelson. 2010. The Economic Cost of Foodborne Disease in New Zealand. Sydney, Australia: Applied Economics.

Galperin, M. Y., D. M. Kristensen, K. S. Makarova, Y. I. Wolf, and E. V. Koonin. 2017. "Microbial genome analysis: the COG approach." *Brief Bioinform*. doi: 10.1093/bib/bbx117.

Galperin, M. Y., K. S. Makarova, Y. I. Wolf, and E. V. Koonin. 2015. "Expanded microbial genome coverage and improved protein family annotation in the COG database." *Nucleic Acids Res* 43 (Database issue):D261-9. doi: 10.1093/nar/gku1223.

Garrity, G. M., J. A. Bell, and T. Lilburn. 2005a. "Class V. *Epsilonproteobabacteria*." In *Bergey's Manual of Systematic Bacteriology*, edited by D. J. Brenner, N. R. Krieg and J. T. Staley, 1145. USA: Springer.

Garrity, G. M., J. A. Bell, and T. Lilburn. 2005b. "Family II. *Helicobacteraceae*." In *Bergey's Manual of Systematic Bacteriology*, edited by D. J. Brenner, N. R. Krieg and J. T. Staley, 1168. USA: Springer.

Garrity, G. M., J. A. Bell, and T. Lilburn. 2005c. "Order I. *Campylobacterales*." In *Bergey's Manual of Systematic Bacteriology*, edited by D. J. Brenner, N. R. Krieg and J. T. Staley, 1145. USA: Springer.

GBD 2015 Disease and Injury Incidence and Prevalence Collaborators. 2016. "Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015." *Lancet* 388 (10053):1545-1602. doi: 10.1016/s0140-6736(16)31678-6.

Geis, G., H. Leying, S. Suerbaum, and W. Opferkuch. 1990. "Unusual fatty acid substitution in lipids and lipopolysaccharides of *Helicobacter pylori*." *J Clin Microbiol* 28 (5):930-2.

Gilbert, M. J., M. Kik, W. G. Miller, B. Duim, and J. A. Wagenaar. 2015. "*Campylobacter iguaniorum* sp. nov., isolated from reptiles." *Int J Syst Evol Microbiol* 65 (Pt 3):975-82. doi: 10.1099/ijs.0.000048.

Gilbert, M. J., W. G. Miller, J. S. Leger, M. H. Chapman, A. J. Timmerman, B. Duim, G. Foster, and J. A. Wagenaar. 2017. "*Campylobacter pinnipediorum* sp. nov., isolated from pinnipeds, comprising *Campylobacter pinnipediorum* subsp. *pinnipediorum* subsp. nov. and *Campylobacter pinnipediorum* subsp. *caledonicus* subsp. nov." *Int J Syst Evol Microbiol* 67 (6):1961-1968. doi: 10.1099/ijsem.0.001894.

Gomez-Camarasa, C., J. Gutierrez-Fernandez, J. M. Rodriguez-Granger, A. Sampedro-Martinez, A. Sorlozano-Puerto, and J. M. Navarro-Mari. 2014. "Evaluation of the rapid RIDAQUICK *Campylobacter*(R) test in a general hospital." *Diagn Microbiol Infect Dis* 78 (2):101-4. doi: 10.1016/j.diagmicrobio.2013.11.009.

Gonzalez, A., Y. Moreno, R. Gonzalez, J. Hernandez, and M. A. Ferrus. 2006. "Development of a simple and rapid method based on polymerase chain reaction-based restriction fragment length polymorphism analysis to differentiate *Helicobacter, Campylobacter*, and *Arcobacter* species." *Curr Microbiol* 53 (5):416-21. doi: 10.1007/s00284-006-0168-5.

Goodwin, S., J. D. McPherson, and W. R. McCombie. 2016. "Coming of age: ten years of next-generation sequencing technologies." *Nat Rev Genet* 17 (6):333-51. doi: 10.1038/nrg.2016.49.

Goossens, H., L. Vlaes, M. De Boeck, B. Pot, K. Kersters, J. Levy, P. De Mol, J. P. Butzler, and P. Vandamme. 1990. "Is "*Campylobacter upsaliensis*" an unrecognised cause of human diarrhoea?" *Lancet* 335 (8689):584-6. doi: 10.1016/0140-6736(90)90359-D.

Goris, J., K. T. Konstantinidis, J. A. Klappenbach, T. Coenye, P. Vandamme, and J. M. Tiedje. 2007. "DNA-DNA hybridization values and their relationship to whole-genome sequence similarities." *Int J Syst Evol Microbiol* 57 (Pt 1):81-91. doi: 10.1099/ijs.0.64483-0.

Gorkiewicz, G., G. Feierl, C. Schober, F. Dieber, J. Kofer, R. Zechner, and E. L. Zechner. 2003. "Species-specific identification of campylobacters by partial 16S rRNA gene sequencing." *J Clin Microbiol* 41 (6):2537-46. doi: 10.1128/jcm.41.6.2537-2546.2003.

Gotz, S., J. M. Garcia-Gomez, J. Terol, T. D. Williams, S. H. Nagaraj, M. J. Nueda, M. Robles, M. Talon, J. Dopazo, and A. Conesa. 2008. "High-throughput functional annotation and data mining with the Blast2GO suite." *Nucleic Acids Res* 36 (10):3420-35. doi: 10.1093/nar/gkn176.

Grange, Z. L., B. D. Gartrell, P. J. Biggs, N. J. Nelson, M. Anderson, and N. P. French. 2016. "Microbial genomics of a host-associated commensal bacterium in fragmented populations of endangered takahe." *Microb Ecol* 71 (4):1020-1029. doi: 10.1007/s00248-015-0721-5.

Greiner, M. 1995. "Two-graph receiver operating characteristic (TG-ROC): a Microsoft-EXCEL template for the selection of cut-off values in diagnostic tests." *J Immunol Methods* 185 (1):145-6. doi: 10.1016/0022-1759(95)00078-O.

Grosche, A., H. Sekaran, I. Perez-Rodriguez, V. Starovoytov, and C. Vetriani. 2015. "*Cetia pacifica* gen. nov., sp. nov., a chemolithoautotrophic, thermophilic, nitrate-ammonifying bacterium from a deep-sea hydrothermal vent." *Int J Syst Evol Microbiol* 65 (Pt 4):1144-50. doi: 10.1099/ijs.0.000070.

Guarino, A., S. Ashkenazi, D. Gendrel, A. Lo Vecchio, R. Shamir, and H. Szajewska. 2014. "European Society for Pediatric Gastroenterology, Hepatology, and Nutrition/European Society for Pediatric Infectious Diseases evidence-based guidelines for the management of acute gastroenteritis in children in Europe: update 2014." *J Pediatr Gastroenterol Nutr* 59 (1):132-52. doi: 10.1097/mpg.0000000000000375.

Guerrant, R. L., and D. A. Bobak. 1991. "Bacterial and protozoal gastroenteritis." *N Engl J Med* 325 (5):327-40. doi: 10.1056/nejm199108013250506.

Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler. 2013. "QUAST: quality assessment tool for genome assemblies." *Bioinformatics* 29 (8):1072-5. doi: 10.1093/bioinformatics/btt086.

Hanninen, M. L., R. I. Karenlampi, J. M. K. Koort, T. Mikkonen, and K. J. Bjorkroth. 2005. "Extension of the species *Helicobacter bilis* to include the reference strains of *Helicobacter* sp. flexispira taxa 2, 3 and 8 and Finnish canine and feline flexispira strains." *Int J Syst Evol Microbiol* 55 (Pt 2):891-8. doi: 10.1099/ijs.0.63245-0.

Hanninen, M. L., M. Utriainen, I. Happonen, and F. E. Dewhirst. 2003. "*Helicobacter* sp. flexispira 16S rDNA taxa 1, 4 and 5 and Finnish porcine *Helicobacter* isolates are

members of the species *Helicobacter trogontum* (taxon 6)." *Int J Syst Evol Microbiol* 53 (Pt 2):425-33. doi: 10.1099/ijs.0.02389-0.

Hansen, R., S. H. Berry, I. Mukhopadhya, J. M. Thomson, K. A. Saunders, C. E. Nicholl, W. M. Bisset, S. Loganathan, G. Mahdi, D. Kastner-Cole, A. R. Barclay, J. Bishop, D. M. Flynn, P. McGrogan, R. K. Russell, E. M. El-Omar, and G. L. Hold. 2013. "The microaerophilic microbiota of *de-novo* paediatric inflammatory bowel disease: the BISCUIT study." *PLoS One* 8 (3):e58825. doi: 10.1371/journal.pone.0058825.

Hansson, I., M. Persson, L. Svensson, E. O. Engvall, and K. E. Johansson. 2008. "Identification of nine sequence types of the 16S rRNA genes of *Campylobacter jejuni* subsp. *jejuni* isolated from broilers." *Acta Vet Scand* 50:10. doi: 10.1186/1751-0147-50-10.

Hardo, P. G., A. Tugnait, F. Hassan, D. A. F. Lynch, A. P. West, N. P. Mapstone, P. Quirke, D. M. Chalmers, M. J. Kowolik, and A. T. R. Axon. 1995. "*Helicobacter pylori* infection and dental care." *Gut* 37 (1):44-6. doi: 10.1136/gut.37.1.44.

Harrington, C. S., and S. L. W. On. 1999. "Extensive 16S rRNA gene sequence diversity in *Campylobacter hyointestinalis* strains: taxonomic and applied implications." *Int J Syst Bacteriol* 49 (Pt 3):1171-5. doi: 10.1099/00207713-49-3-1171.

Hatanaka, N., A. Shimizu, S. Somroop, Y. Li, M. Asakura, A. Nagita, S. Prasad Awasthi, A. Hinenoya, and S. Yamasaki. 2017. "High prevalence of *Campylobacter ureolyticus* in stool specimens of children with diarrhea in Japan." *Jpn J Infect Dis* 70 (4):455-457. doi: 10.7883/yoken.JJID.2016.428.

Health Intelligence Team. 2014. Notifiable and Other Diseases in New Zealand: Annual Report 2013. Porirua, New Zealand: Institute of Environmental Science and Research.

Health Intelligence Team. 2015. Notifiable Diseases in New Zealand 2014. Porirua, New Zealand: Institute of Environmental Science and Research.

Health Intelligence Team. 2016. Notifiable Diseases in New Zealand 2015. Porirua, New Zealand: Institute of Environmental Science and Research.

Health Intelligence Team. 2017. Notifiable Diseases in New Zealand Annual Report 2016. Porirua, New Zealand: Institute of Environmental Science and Research.

Houf, K., S. L. W. On, T. Coenye, L. Debruyne, S. De Smet, and P. Vandamme. 2009. "*Arcobacter thereius* sp. nov., isolated from pigs and ducks." *Int J Syst Evol Microbiol* 59 (Pt 10):2599-604. doi: 10.1099/ijs.0.006650-0.

Huang, Y. J., and H. A. Boushey. 2015. "The microbiome in asthma." *J Allergy Clin Immunol* 135 (1):25-30. doi: 10.1016/j.jaci.2014.11.011.

Huerta-Cepas, J., K. Forslund, L. P. Coelho, D. Szklarczyk, L. J. Jensen, C. von Mering, and P. Bork. 2017. "Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper." *Mol Biol Evol* 34 (8):2115-2122. doi: 10.1093/molbev/msx148.

Huerta-Cepas, J., D. Szklarczyk, K. Forslund, H. Cook, D. Heller, M. C. Walter, T. Rattei, D. R. Mende, S. Sunagawa, M. Kuhn, L. J. Jensen, C. von Mering, and P. Bork. 2016. "eggNOG 4.5: a hierarchical orthology framework with improved functional annotations

for eukaryotic, prokaryotic and viral sequences." *Nucleic Acids Res* 44 (D1):D286-93. doi: 10.1093/nar/gkv1248.

Huq, M., G. Gonis, and T. Istivan. 2014. "Development and evaluation of a multiplex PCR for the detection of *Campylobacter concisus* and other *Campylobacter* spp. from gastroenteritisi cases." *Open J Med Microbiol* 4:29-37.

Huq, M., T. T. H. Van, V. Gurtler, E. Elshagmani, K. S. Allemailem, P. M. Smooker, and T. Istivan. 2017. "The ribosomal RNA operon (*rrn*) of *Campylobacter concisus* supports molecular typing to genomospecies level." *Gene Reports* 6:8-14. doi: 10.1016/j.genrep.2016.10.008.

Huson, D. H., and D. Bryant. 2006. "Application of phylogenetic networks in evolutionary studies." *Mol Biol Evol* 23 (2):254-67. doi: 10.1093/molbev/msj030.

Hyatt, D., G. L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser. 2010. "Prodigal: prokaryotic gene recognition and translation initiation site identification." *BMC Bioinformatics* 11:119. doi: 10.1186/1471-2105-11-119.

Ibrahim, A., B. M. Goebel, W. Liesack, M. Griffiths, and E. Stackebrandt. 1993. "The phylogeny of the genus *Yersinia* based on 16S rDNA sequences." *FEMS Microbiol Lett* 114 (2):173-7. doi: 10.1111/j.1574-6968.1993.tb06569.x.

Inagaki, F., K. Takai, H. Kobayashi, K. H. Nealson, and K. Horikoshi. 2003. "*Sulfurimonas autotrophica* gen. nov., sp. nov., a novel sulfur-oxidizing epsilon-proteobacterium isolated from hydrothermal sediments in the Mid-Okinawa Trough." *Int J Syst Evol Microbiol* 53 (Pt 6):1801-5. doi: 10.1099/ijs.0.02682-0.

Inagaki, F., K. Takai, K. H. Nealson, and K. Horikoshi. 2004. "*Sulfurovum lithotrophicum* gen. nov., sp. nov., a novel sulfur-oxidizing chemolithoautotroph within the *epsilon-Proteobacteria* isolated from Okinawa Trough hydrothermal sediments." *Int J Syst Evol Microbiol* 54 (Pt 5):1477-82. doi: 10.1099/ijs.0.03042-0.

Inglis, G. D., V. F. Boras, and A. Houde. 2011. "Enteric campylobacteria and RNA viruses associated with healthy and diarrheic humans in the Chinook health region of southwestern Alberta, Canada." *J Clin Microbiol* 49 (1):209-19. doi: 10.1128/jcm.01220-10.

Inglis, G. D., and L. D. Kalischuk. 2003. "Use of PCR for direct detection of *Campylobacter* species in bovine feces." *Appl Environ Microbiol* 69 (6):3435-47. doi: 10.1128/aem.69.6.3435–3447.2003.

Intenational Commission on Microbiological specifications for Foods. 2002. "Ranking of foodborne pathogens of toxins into hazard groups." In *Microorganisms in Foods 7*, edited by R. B. Tompkin, 166-172. New York, N.Y.: Kluwer Academic/Plenum Publishers.

Ismail, Y., V. Mahendran, S. Octavia, A. S. Day, S. M. Riordan, M. C. Grimm, R. Lan, D. Lemberg, T. A. T. Tran, and L. Zhang. 2012. "Investigation of the enteric pathogenic potential of oral *Campylobacter concisus* strains isolated from patients with

inflammatory bowel disease." *PLoS One* 7 (5):e38217. doi: 10.1371/journal.pone.0038217.

Istivan, T. S., P. J. Coloe, B. N. Fry, P. Ward, and S. C. Smith. 2004. "Characterization of a haemolytic phospholipase A_2 activity in clinical isolates of *Campylobacter concisus*." *J Med Microbiol* 53 (Pt 6):483-93. doi: 10.1099/jmm.0.45554-0.

Jani, I. V., G. Janossy, D. W. G. Brown, and F. Mandy. 2002. "Multiplexed immunoassays by flow cytometry for diagnosis and surveillance of infectious diseases in resource-poor settings." *Lancet Infect Dis* 2 (4):243-50. doi: 10.1016/S1473-3099(02)00242-6.

Jensen, L. J., P. Julien, M. Kuhn, C. von Mering, J. Muller, T. Doerks, and P. Bork. 2008. "eggNOG: automated construction and annotation of orthologous groups of genes." *Nucleic Acids Res* 36 (Database issue):D250-4. doi: 10.1093/nar/gkm796.

Jimenez, G., M. Urdiain, A. Cifuentes, A. Lopez-Lopez, A. R. Blanch, J. Tamames, P. Kampfer, A. B. Kolsto, D. Ramon, J. F. Martinez, F. M. Codoner, and R. Rossello-Mora. 2013. "Description of *Bacillus toyonensis* sp. nov., a novel species of the *Bacillus cereus* group, and pairwise genome comparisons of the species of the group by means of ANI calculations." *Syst Appl Microbiol* 36 (6):383-91. doi: 10.1016/j.syapm.2013.04.008.

Jolley, K. A., C. M. Bliss, J. S. Bennett, H. B. Bratcher, C. Brehony, F. M. Colles, H. Wimalarathna, O. B. Harrison, S. K. Sheppard, A. J. Cody, and M. C. Maiden. 2012. "Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain." *Microbiology* 158 (Pt 4):1005-15. doi: 10.1099/mic.0.055459-0.

Jones, T. F., and P. Gerner-Smidt. 2012. "Nonculture diagnostic tests for enteric diseases." *Emerg Infect Dis* 18 (3):513-4. doi: 10.3201/eid1803.111914.

Kaakoush, N. O., N. Castano-Rodriguez, A. S. Day, D. A. Lemberg, S. T. Leach, and H. M. Mitchell. 2014. "*Campylobacter concisus* and exotoxin 9 levels in paediatric patients with Crohn's disease and their association with the intestinal microbiota." *J Med Microbiol* 63 (Pt 1):99-105. doi: 10.1099/jmm.0.067231-0.

Kaakoush, N. O., N. P. Deshpande, M. R. Wilkins, C. G. Tan, J. A. Burgos-Portugal, M. J. Raftery, A. S. Day, D. A. Lemberg, and H. Mitchell. 2011. "The pathogenic potential of *Campylobacter concisus* strains associated with chronic intestinal diseases." *PLoS One* 6 (12):e29045. doi: 10.1371/journal.pone.0029045.

Kaakoush, N. O., N. Sodhi, J. W. Chenu, J. M. Cox, S. M. Riordan, and H. M. Mitchell. 2014. "The interplay between *Campylobacter* and *Helicobacter* species and other gastrointestinal microbiota of commercial broiler chickens." *Gut Pathog* 6:18. doi: 10.1186/1757-4749-6-18.

Kalischuk, L. D., and G. D. Inglis. 2011. "Comparative genotypic and pathogenic examination of *Campylobacter concisus* isolates from diarrheic and non-diarrheic humans." *BMC Microbiol* 11:53. doi: 10.1186/1471-2180-11-53.

Kamei, K., H. Kawabata, M. Asakura, W. Samosornsuk, A. Hinenoya, S. Nakagawa, and S. Yamasaki. 2016. "A cytolethal distending toxin gene-based multiplex PCR assay for

Campylobacter jejuni, C. fetus, C. coli, C. upsaliensis, C. hyointestinalis, and *C. lari*." *Jpn J Infect Dis* 69 (3):256-8. doi: 10.7883/yoken.JJID.2015.182.

Kasper, M. R., A. G. Lescano, C. Lucas, D. Gilles, B. J. Biese, G. Stolovitz, and E. J. Reaves. 2012. "Diarrhea outbreak during U.S. military training in El Salvador." *PLoS One* 7 (7):e40404. doi: 10.1371/journal.pone.0040404.

Kaur, T., J. Singh, M. A. Huffman, K. J. Petrzelkova, N. S. Taylor, S. Xu, F. E. Dewhirst, B. J. Paster, L. Debruyne, P. Vandamme, and J. G. Fox. 2011. "*Campylobacter troglodytis* sp. nov., isolated from feces of human-habituated wild chimpanzees (*Pan troglodytes schweinfurthii*) in Tanzania." *Appl Environ Microbiol* 77 (7):2366-73. doi: 10.1128/aem.01840-09.

Kawasaki, S., P. M. Fratamico, I. V. Wesley, and S. Kawamoto. 2008. "Species-specific identification of campylobacters by PCR-restriction fragment length polymorphism and PCR targeting of the gyrase B gene." *Appl Environ Microbiol* 74 (8):2529-33. doi: 10.1128/aem.00975-07.

Keenan, J. I., J. D. Leaman, R. Greenlees, and A. S. Day. 2014. "Is *Campylobacter consisus* an unrecognised cause of diarrhoea in New Zealand?" *N Z Med J* 127 (1391):90-1.

Kersey, P. J., J. E. Allen, I. Armean, S. Boddu, B. J. Bolt, D. Carvalho-Silva, M. Christensen, P. Davis, L. J. Falin, C. Grabmueller, J. Humphrey, A. Kerhornou, J. Khobova, N. K. Aranganathan, N. Langridge, E. Lowy, M. D. McDowall, U. Maheswari, M. Nuhn, C. K. Ong, B. Overduin, M. Paulini, H. Pedro, E. Perry, G. Spudich, E. Tapanari, B. Walts, G. Williams, M. Tello-Ruiz, J. Stein, S. Wei, D. Ware, D. M. Bolser, K. L. Howe, E. Kulesha, D. Lawson, G. Maslen, and D. M. Staines. 2016. "Ensembl Genomes 2016: more genomes, more complexity." *Nucleic Acids Res* 44 (D1):D574-80. doi: 10.1093/nar/gkv1209.

Kim, H. M., C. Y. Hwang, and B. C. Cho. 2010. "*Arcobacter marinus* sp. nov." *Int J Syst Evol Microbiol* 60 (Pt 3):531-6. doi: 10.1099/ijs.0.007740-0.

Kim, S.-Y., B. Chung, J. H. Chang, G. Y. Jung, H. W. Kim, B. Y. Park, S. S. Oh, and M. H. Oh. 2016. "Simultaneous identification of 13 foodborne pathogens by using capillary electrophoresis-single strand conformation polymorphism coupled with multiplex ligation-dependent probe amplification and its application in foods." *Foodborne Pathog Dis* 13 (10):566-574. doi: 10.1089/fpd.2016.2143.

King, C. K., R. Glass, J. S. Bresee, and C. Duggan. 2003. "Managing acute gastroenteritis among children: oral rehydration, maintenance, and nutritional therapy." *MMWR Recomm Rep* 52 (RR-16):1-16.

Klein, E. J., D. R. Boster, J. R. Stapp, J. G. Wells, X. Qin, C. R. Clausen, D. L. Swerdlow, C. R. Braden, and P. I. Tarr. 2006. "Diarrhea etiology in a children's hospital emergency department: a prospective cohort study." *Clin Infect Dis* 43 (7):807-13. doi: 10.1086/507335.

Klena, J. D., C. T. Parker, K. Knibb, J. C. Ibbitt, P. M. L. Devane, S. T. Horn, W. G. Miller, and M. E. Konkel. 2004. "Differentiation of *Campylobacter coli*, *Campylobacter jejuni*,

Campylobacter lari, and *Campylobacter upsaliensis* by a multiplex PCR developed from the nucleotide sequence of the lipid A gene *lpx*A." *J Clin Microbiol* 42 (12):5549-57. doi: 10.1128/jcm.42.12.5549-5557.2004.

Klimke, W., R. Agarwala, A. Badretdin, S. Chetvernin, S. Ciufo, B. Fedorov, B. Kiryutin, K. O'Neill, W. Resch, S. Resenchuk, S. Schafer, I. Tolstoy, and T. Tatusova. 2009. "The National Center for Biotechnology Information's Protein Clusters Database." *Nucleic Acids Res* 37 (Database issue):D216-23. doi: 10.1093/nar/gkn734.

Kodama, Y., L. T. Ha, and K. Watanabe. 2007. "*Sulfurospirillum cavolei* sp. nov., a facultatively anaerobic sulfur-reducing bacterium isolated from an underground crude oil storage cavity." *Int J Syst Evol Microbiol* 57 (Pt 4):827-31. doi: 10.1099/ijs.0.64823-0.

Kodama, Y., and K. Watanabe. 2004. "*Sulfuricurvum kujiense* gen. nov., sp. nov., a facultatively anaerobic, chemolithoautotrophic, sulfur-oxidizing bacterium isolated from an underground crude-oil storage cavity." *Int J Syst Evol Microbiol* 54 (Pt 6):2297-300. doi: 10.1099/ijs.0.63243-0.

Kohn, M. A., C. R. Carpenter, and T. B. Newman. 2013. "Understanding the direction of bias in studies of diagnostic test accuracy." *Acad Emerg Med* 20 (11):1194-206. doi: 10.1111/acem.12255.

Kondrashova, O., C. J. Love, S. Lunke, A. L. Hsu, Australian Ovarian Cancer Study (AOCS) Group, P. M. Waring, and G. R. Taylor. 2015. "High-throughput amplicon-based copy number detection of 11 genes in formalin-fixed paraffin-embedded ovarian tumour samples by MLPA-Seq." *PLoS One* 10 (11):e0143006. doi: 10.1371/journal.pone.0143006.

Konstantinidis, K. T., A. Ramette, and J. M. Tiedje. 2006. "Toward a more robust assessment of intraspecies diversity, using fewer genetic markers." *Appl Environ Microbiol* 72 (11):7286-93. doi: 10.1128/aem.01398-06.

Konstantinidis, K. T., and J. M. Tiedje. 2005. "Genomic insights that advance the species definition for prokaryotes." *Proc Natl Acad Sci U S A* 102 (7):2567-72. doi: 10.1073/pnas.0409727102.

Korczak, B. M., R. Stieber, S. Emler, A. P. Burnens, J. Frey, and P. Kuhnert. 2006. "Genetic relatedness within the genus *Campylobacter* inferred from *rpo*B sequences." *Int J Syst Evol Microbiol* 56 (Pt 5):937-45. doi: 10.1099/ijs.0.64109-0.

Kotetishvili, M., A. Kreger, G. Wauters, J. G. Morris, Jr., A. Sulakvelidze, and O. C. Stine. 2005. "Multilocus sequence typing for studying genetic relationships among *Yersinia* species." *J Clin Microbiol* 43 (6):2674-84. doi: 10.1128/jcm.43.6.2674-2684.2005.

Koziel, M., D. Corcoran, I. O'Callaghan, R. D. Sleator, and B. Lucey. 2013. "Validation of the EntericBio Panel II(R) multiplex polymerase chain reaction system for detection of *Campylobacter* spp., *Salmonella* spp., *Shigella* spp., and verotoxigenic *E. coli* for use in a clinical diagnostic setting." *Diagn Microbiol Infect Dis* 75 (1):46-9. doi: 10.1016/j.diagmicrobio.2012.09.007.

Koziel, M., B. Lucey, S. Bullman, G. D. Corcoran, and R. D. Sleator. 2012. "Molecular-based detection of the gastrointestinal pathogen *Campylobacter ureolyticus* in unpasteurized milk samples from two cattle farms in Ireland." *Gut Pathog* 4 (1):14. doi: 10.1186/1757-4749-4-14.

Koziel, M., P. O'Doherty, P. Vandamme, G. D. Corcoran, R. D. Sleator, and B. Lucey. 2014. "*Campylobacter corcagiensis* sp. nov., isolated from faeces of captive lion-tailed macaques (*Macaca silenus*)." *Int J Syst Evol Microbiol* 64 (Pt 8):2878-83. doi: 10.1099/ijs.0.063867-0.

Kreader, C. A. 1996. "Relief of amplification inhibition in PCR with bovine serum albumin or T4 gene 32 protein." *Appl Environ Microbiol* 62 (3):1102-6.

Kriventseva, E. V., F. Tegenfeldt, T. J. Petty, R. M. Waterhouse, F. A. Simao, I. A. Pozdnyakov, P. Ioannidis, and E. M. Zdobnov. 2015. "OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software." *Nucleic Acids Res* 43 (Database issue):D250-6. doi: 10.1093/nar/gku1220.

Kroger, A., O. Klimmek, P. Vandamme, F. E. Dewhirst, and B. J. Paster. 2005. "Genus III. *Wolinella*." In *Bergey's Manual of Systematic Bacteriology*, edited by D. J. Brenner, N. R. Krieg and J. T Staley, 1191-4. USA: Springer.

Labrenz, M., J. Grote, K. Mammitzsch, H. T. S. Boschker, M. Laue, G. Jost, S. Glaubitz, and K. Jurgens. 2013. "*Sulfurimonas gotlandica* sp. nov., a chemoautotrophic and psychrotolerant epsilonproteobacterium isolated from a pelagic redoxcline, and an emended description of the genus *Sulfurimonas*." *Int J Syst Evol Microbiol* 63 (Pt 11):4141-8. doi: 10.1099/ijs.0.048827-0.

Laing, C., C. Buchanan, E. N. Taboada, Y. Zhang, A. Kropinski, A. Villegas, J. E. Thomas, and V. P. J. Gannon. 2010. "Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions." *BMC Bioinformatics* 11:461. doi: 10.1186/1471-2105-11-461.

Lake, R. J., S. B. Adlam, S. Perera, D. M. Campbell, and M. G. Baker. 2010. "The disease pyramid for acute gastrointestinal illness in New Zealand." *Epidemiol Infect* 138 (10):1468-71. doi: 10.1017/s0950268810000397.

Lake, R. J., N. King, K. Sexton, P. Bridgewater, and D. Campbell. 2009. "Acute gastrointestinal illness in New Zealand: information from a survey of community and hospital laboratories." *N Z Med J* 122 (1307):48-54.

Land, M., L. Hauser, S. R. Jun, I. Nookaew, M. R. Leuze, T. H. Ahn, T. Karpinets, O. Lund, G. Kora, T. Wassenaar, S. Poudel, and D. W. Ussery. 2015. "Insights from 20 years of bacterial genome sequencing." *Funct Integr Genomics* 15 (2):141-61. doi: 10.1007/s10142-015-0433-4.

Lanza, V. F., F. Baquero, F. de la Cruz, and T. M. Coque. 2017. "AcCNET (Accessory Genome Constellation Network): comparative genomics software for accessory genome analysis using bipartite networks." *Bioinformatics* 33 (2):283-285. doi: 10.1093/bioinformatics/btw601.

Lapage, S. P., P. H. A. Sneath, E. F. Lessel, V. B. D. Skerman, H. P. R. Seeliger, and W. A. Clark, eds. 1992. *International Code of Nomenclature of Bacteria: Bacteriological Code, 1990 Revision*. Washington (DC): ASM Press and International Union of Microbiological Societies.

Larsen, M. V., S. Cosentino, O. Lukjancenko, D. Saputra, S. Rasmussen, H. Hasman, T. Sicheritz-Ponten, F. M. Aarestrup, D. W. Ussery, and O. Lund. 2014. "Benchmarking of methods for genomic taxonomy." *J Clin Microbiol* 52 (5):1529-39. doi: 10.1128/jcm.02981-13.

Lastovica, A. J. 2006. "Emerging *Campylobacter* spp.: the tip of the iceberg." *Clin Microbiol News* 28 (7):49-56.

Lastovica, A. J., and E. Le Roux. 2000. "Efficient isolation of campylobacteria from stools." *J Clin Microbiol* 38 (7):2798-9.

Lastovica, A. J., E. Le Roux, R. Warren, and H. Klump. 1994. "Additional data on clinical isolates of *Campylobacter mucosalis*." *J Clin Microbiol* 32 (9):2338-9.

Lastovica, A. J., S. L. W. On, and L. Zhang. 2014. "The family *Campylobacteraceae*." In *The Prokaryotes - Deltaproteobacteria and Epsilonproteobacteria*, edited by E. Rosenberg, E. F. Delong, S. Lory, E. Stackebrandt and F. Thompson, 307-335. Berlin Heidelberg: Springer-Verlag.

Lastovica, A. J., and M.B. Skirrow. 2000. "Clinical significance of *Campylobacter* and related species other than *Campylobacter jejuni* and *C. coli*." In *Campylobacter*, edited by I. Nachamkin and M. J. Blaser, 89-120. Washington, D.C.: ASM Press.

Lastovica, A., E. Le Roux, R. Warren, and H. Klump. 1993. "Clinical isolates of *Campylobacter mucosalis*." *J Clin Microbiol* 31 (10):2835-6.

Lausch, K. R., L. Westh, L. H. Kristensen, J. Lindberg, B. Tarp, and C. S. Larsen. 2017. "Rotavirus is frequent among adults hospitalised for acute gastroenteritis." *Dan Med J* 64 (1):A5312.

Lawson, A. J., S. L. W. On, J. M. J. Logan, and J. Stanley. 2001. "*Campylobacter hominis* sp. nov., from the human gastrointestinal tract." *Int J Syst Evol Microbiol* 51 (Pt 2):651-60. doi: 10.1099/00207713-51-2-651.

Lawson, G. H. K., J. L. Leaver, G. W. Pettigrew, and A. C. Rowland. 1981. "Some features of *Campylobacter sputorum* subsp. *mucosalis* subsp. nov., nom. rev. and their taxonomic significance." *Int J Syst Bacteriol* 31 (4):385-391. doi: 10.1099/00207713-31-4-385.

Lee, C., C. Grasso, and M. F. Sharlow. 2002. "Multiple sequence alignment using partial order graphs." *Bioinformatics* 18 (3):452-64. doi: 10.1093/bioinformatics/18.3.452.

Lehninger, A. L. 1982. *Principles of Biochemistry*. New York, USA: Worth Publishers.

Lequin, R. M. 2005. "Enzyme immunoassay (EIA)/enzyme-linked immunosorbent assay (ELISA)." *Clin Chem* 51 (12):2415-8. doi: 10.1373/clinchem.2005.051532.

Levican, A., L. Collado, C. Aguilar, C. Yustes, A. L. Dieguez, J. L. Romalde, and M. J. Figueras. 2012. "*Arcobacter bivalviorum* sp. nov. and *Arcobacter venerupis* sp. nov., new species

isolated from shellfish." *Syst Appl Microbiol* 35 (3):133-8. doi: 10.1016/j.syapm.2012.01.002.

Levican, A., L. Collado, and M. J. Figueras. 2013. "*Arcobacter cloacae* sp. nov. and *Arcobacter suis* sp. nov., two new species isolated from food and sewage." *Syst Appl Microbiol* 36 (1):22-7. doi: 10.1016/j.syapm.2012.11.003.

Levican, A., S. Rubio-Arcos, A. Martinez-Murcia, L. Collado, and M. J. Figueras. 2015. "*Arcobacter ebronensis* sp. nov. and *Arcobacter aquimarinus* sp. nov., two new species isolated from marine environment." *Syst Appl Microbiol* 38 (1):30-5. doi: 10.1016/j.syapm.2014.10.011.

Li, L., C. J. Stoeckert, Jr., and D. S. Roos. 2003. "OrthoMCL: identification of ortholog groups for eukaryotic genomes." *Genome Res* 13 (9):2178-89. doi: 10.1101/gr.1224503.

Lindblom, G. B., E. Sjogren, J. Hansson-Westerberg, and B. Kaijser. 1995. "*Campylobacter upsaliensis, C. sputorum sputorum* and *C. concisus* as common causes of diarrhoea in Swedish children." *Scand J Infect Dis* 27 (2):187-8.

Linton, D., R. J. Owen, and J. Stanley. 1996. "Rapid identification by PCR of the genus *Campylobacter* and of five *Campylobacter* species enteropathogenic for man and animals." *Res Microbiol* 147 (9):707-18. doi: 10.1016/S0923-2508(97)85118-2.

Lipkin, W. I. 2010. "Microbe hunting." *Microbiol Mol Biol Rev* 74 (3):363-77. doi: 10.1128/mmbr.00007-10.

Liu, J., F. Kabir, J. Manneh, P. Lertsethtakarn, S. Begum, J. Gratz, S. M. Becker, D. J. Operario, M. Taniuchi, L. Janaki, J. A. Platts-Mills, D. M. Haverstick, M. Kabir, S. U. Sobuz, K. Nakjarung, P. Sakpaisal, S. Silapong, L. Bodhidatta, S. Qureshi, A. Kalam, Q. Saidi, N. Swai, B. Mujaga, A. Maro, B. Kwambana, M. Dione, M. Antonio, G. Kibiki, C. J. Mason, R. Haque, N. Iqbal, A. K. M. Zaidi, and E. R. Houpt. 2014. "Development and assessment of molecular diagnostic tests for 15 enteropathogens causing childhood diarrhoea: a multicentre study." *Lancet Infect Dis* 14 (8):716-724. doi: 10.1016/s1473-3099(14)70808-4.

Liu, L., Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law. 2012. "Comparison of next-generation sequencing systems." *J Biomed Biotechnol* 2012:251364. doi: 10.1155/2012/251364.

Lopez, L., F. J. Castillo, A. Clavel, and M. C. Rubio. 1998. "Use of a selective medium and a membrane filter method for isolation of *Campylobacter* species from Spanish paediatric patients." *Eur J Clin Microbiol Infect Dis* 17 (7):489-92.

Lorrot, M., F. Bon, M. J. El Hajje, S. Aho, M. Wolfer, H. Giraudon, J. Kaplon, E. Marc, J. Raymond, P. Lebon, P. Pothier, and D. Gendrel. 2011. "Epidemiology and clinical features of gastroenteritis in hospitalised children: prospective survey during a 2-year period in a Parisian hospital, France." *Eur J Clin Microbiol Infect Dis* 30 (3):361-8. doi: 10.1007/s10096-010-1094-9.

Luijten, M. L. G. C, J. de Weert, H. Smidt, H. T. S. Boschker, W. M. de Vos, G. Schraa, and A. J. M. Stams. 2003. "Description of *Sulfurospirillum halorespirans* sp. nov., an anaerobic,

tetrachloroethene-respiring bacterium, and transfer of *Dehalospirillum multivorans* to the genus *Sulfurospirillum* as *Sulfurospirillum multivorans* comb. nov." *Int J Syst Evol Microbiol* 53 (Pt 3):787-93. doi: 10.1099/ijs.0.02417-0.

Lukashin, A. V., and M. Borodovsky. 1998. "GeneMark.hmm: new solutions for gene finding." *Nucleic Acids Res* 26 (4):1107-15.

Lukjancenko, O, MC Thomsen, M Voldby Larsen, and DW Ussery. 2013. "PanFunPro: PAN-genome analysis based on FUNctional PROfiles." *F1000Res* 2:265.

Lukjancenko, O., D. W. Ussery, and T. M. Wassenaar. 2012. "Comparative genomics of *Bifidobacterium*, *Lactobacillus* and related probiotic genera." *Microb Ecol* 63 (3):651-73. doi: 10.1007/s00248-011-9948-y.

Lukjancenko, O., T. M. Wassenaar, and D. W. Ussery. 2010. "Comparison of 61 sequenced *Escherichia coli* genomes." *Microb Ecol* 60 (4):708-20. doi: 10.1007/s00248-010-9717-3.

Lynch, O. A., C. Cagney, D. A. McDowell, and G. Duffy. 2011. "Occurrence of fastidious *Campylobacter* spp. in fresh meat and poultry using an adapted cultural protocol." *Int J Food Microbiol* 150 (2-3):171-7. doi: 10.1016/j.ijfoodmicro.2011.07.037.

Lynch, T., A. Petkau, N. Knox, M. Graham, and G. Van Domselaar. 2016. "A primer on infectious disease bacterial genomics." *Clin Microbiol Rev* 29 (4):881-913. doi: 10.1128/cmr.00001-16.

M'Ikanatha N, M., L. A. Dettinger, A. Perry, P. Rogers, S. M. Reynolds, and I. Nachamkin. 2012. "Culturing stool specimens for *Campylobacter* spp., Pennsylvania, USA." *Emerg Infect Dis* 18 (3):484-7. doi: 10.3201/eid1803.111266.

Macuch, P. J., and A. C. R. Tanner. 2000. "*Campylobacter* species in health, gingivitis, and periodontitis." *J Dent Res* 79 (2):785-92. doi: 10.1177/00220345000790021301.

Maddison, D. R., D. L. Swofford, and W. P. Maddison. 1997. "NEXUS: an extensible file format for systematic information." *Syst Biol* 46 (4):590-621.

Maertens de Noordhout, C., B. Devleesschauwer, J. A. Haagsma, A. H. Havelaar, S. Bertrand, O. Vandenberg, S. Quoilin, P. T. Brandt, and N. Speybroeck. 2017. "Burden of salmonellosis, campylobacteriosis and listeriosis: a time series analysis, Belgium, 2012 to 2020." *Euro Surveill* 22 (38). doi: 10.2807/1560-7917.es.2017.22.38.30615.

Magoc, T., S. Pabinger, S. Canzar, X. Liu, Q. Su, D. Puiu, L. J. Tallon, and S. L. Salzberg. 2013. "GAGE-B: an evaluation of genome assemblers for bacterial organisms." *Bioinformatics* 29 (14):1718-25. doi: 10.1093/bioinformatics/btt273.

Mahendran, V., S. Octavia, O. Faruk Demirbas, S. Sabrina, R. Ma, R. Lan, M. Riordan S, M. C. Grimm, and L. Zhang. 2015. "Delineation of genetic relatedness and population structure of oral and enteric *Campylobacter concisus* strains by analysis of housekeeping genes." *Microbiology* 161 (8):1600-12. doi: 10.1099/mic.0.000112.

Mahendran, V., S. M. Riordan, M. C. Grimm, T. A. T. Tran, J. Major, N. O. Kaakoush, H. Mitchell, and L. Zhang. 2011. "Prevalence of *Campylobacter* species in adult Crohn's

disease and the preferential colonization sites of *Campylobacter* species in the human intestine." *PLoS One* 6 (9):e25417. doi: 10.1371/journal.pone.0025417.

Mahendran, V., Y. S. Tan, S. M. Riordan, M. C. Grimm, A. S. Day, D. A. Lemberg, S. Octavia, R. Lan, and L. Zhang. 2013. "The prevalence and polymorphisms of zonula occluden toxin gene in multiple *Campylobacter concisus* strains isolated from saliva of patients with inflammatory bowel disease and controls." *PLoS One* 8 (9):e75525. doi: 10.1371/journal.pone.0075525.

Man, S. M., L. Zhang, A. S. Day, S. T. Leach, D. A. Lemberg, and H. Mitchell. 2010. "*Campylobacter concisus* and other *Campylobacter* species in children with newly diagnosed Crohn's disease." *Inflamm Bowel Dis* 16 (6):1008-16. doi: 10.1002/ibd.21157.

Marder, E. P., P. R. Cieslak, A. B. Cronquist, J. Dunn, S. Lathrop, T. Rabatsky-Ehr, P. Ryan, K. Smith, M. Tobin-D'Angelo, D. J. Vugia, S. Zansky, K. G. Holt, B. J. Wolpert, M. Lynch, R. Tauxe, and A. L. Geissler. 2017. "Incidence and trends of infections with pathogens transmitted commonly through food and the effect of increasing use of culture-independent diagnostic tests on surveillance - Foodborne Diseases Active Surveillance Network, 10 U.S. sites, 2013-2016." *MMWR Morb Mortal Wkly Rep* 66 (15):397-403. doi: 10.15585/mmwr.mm6615a1.

Mardis, E. R. 2013. "Next-generation sequencing platforms." *Annu Rev Anal Chem (Palo Alto Calif)* 6:287-303. doi: 10.1146/annurev-anchem-062012-092628.

Markowitz, V. M., K. Mavromatis, N. N. Ivanova, I. M. A. Chen, K. Chu, and N. C. Kyrpides. 2009. "IMG ER: a system for microbial genome annotation expert review and curation." *Bioinformatics* 25 (17):2271-8. doi: 10.1093/bioinformatics/btp393.

Marshall, S. M., P. L. Melito, D. L. Woodward, W. M. Johnson, F. G. Rodgers, and M. R. Mulvey. 1999. "Rapid identification of *Campylobacter, Arcobacter*, and *Helicobacter* isolates by PCR-restriction fragment length polymorphism analysis of the 16S rRNA gene." *J Clin Microbiol* 37 (12):4158-60.

Martiny, D., A. Dediste, L. Debruyne, L. Vlaes, N. B. Haddou, P. Vandamme, and O. Vandenberg. 2011. "Accuracy of the API Campy system, the Vitek 2 *Neisseria-Haemophilus* card and matrix-assisted laser desorption ionization time-of-flight mass spectrometry for the identification of *Campylobacter* and related organisms." *Clin Microbiol Infect* 17 (7):1001-6. doi: 10.1111/j.1469-0691.2010.03328.x.

Matsheka, M. I., B. G. Elisha, A. L. Lastovica, and S. L. W. On. 2002. "Genetic heterogeneity of *Campylobacter concisus* determined by pulsed field gel electrophoresis-based macrorestriction profiling." *FEMS Microbiol Lett* 211 (1):17-22. doi: 10.1111/j.1574-6968.2002.tb11197.x

Matsheka, M. I., A. J. Lastovica, and B. G. Elisha. 2001. "Molecular identification of *Campylobacter concisus*." *J Clin Microbiol* 39 (10):3684-9. doi: 10.1128/jcm.39.10.3684-3689.2001.

McAuliffe, G. N., T. P. Anderson, M. Stevens, J. Adams, R. Coleman, P. Mahagamasekera, S. Young, T. Henderson, M. Hofmann, L. C. Jennings, and D. R. Murdoch. 2013. "Systematic application of multiplex PCR enhances the detection of bacteria, parasites, and viruses in stool samples." *J Infect* 67 (2):122-9. doi: 10.1016/j.jinf.2013.04.009.

Meier-Kolthoff, J. P., A. F. Auch, H. P. Klenk, and M. Goker. 2013. "Genome sequence-based species delimitation with confidence intervals and improved distance functions." *BMC Bioinformatics* 14:60. doi: 10.1186/1471-2105-14-60.

Meier-Kolthoff, J. P., H. P. Klenk, and M. Goker. 2014. "Taxonomic use of DNA G+C content and DNA-DNA hybridization in the genomic age." *Int J Syst Evol Microbiol* 64 (Pt 2):352-6. doi: 10.1099/ijs.0.056994-0.

Melito, P. L., C. Munro, P. R. Chipman, D. L. Woodward, T. F. Booth, and F. G. Rodgers. 2001. "*Helicobacter winghamensis* sp. nov., a novel *Helicobacter* sp. isolated from patients with gastroenteritis." *J Clin Microbiol* 39 (7):2412-7. doi: 10.1128/jcm.39.7.2412-2417.2001.

Melito, P. L., D. L. Woodward, K. A. Bernard, L. Price, R. Khakhria, W. M. Johnson, and F. G. Rodgers. 2000. "Differentiation of clinical *Helicobacter pullorum* isolates from related *Helicobacter* and *Campylobacter* species." *Helicobacter* 5 (3):142-7. doi: 10.1046/j.1523-5378.2000.00022.x.

Miller, W. G. 2017. "RE: Inquiry-error: Campylobacter concisus Lasto61.99 (SUB1340221 ; BioSample SAMN03457017) ", 14th March 2017.

Miller, W. G., M. H. Chapman, E. Yee, S. L. W. On, D. K. McNulty, A. J. Lastovica, A. M. Carroll, E. B. McNamara, G. Duffy, and R. E. Mandrell. 2012. "Multilocus sequence typing methods for the emerging *Campylobacter* species *C. hyointestinalis, C. lanienae, C. sputorum, C. concisus*, and *C. curvus*." *Front Cell Infect Microbiol* 2:45. doi: 10.3389/fcimb.2012.00045.

Miller, W. G., E. Yee, M. H. Chapman, T. P. L. Smith, J. L. Bono, S. Huynh, C. T. Parker, P. Vandamme, K. Luong, and J. Korlach. 2014. "Comparative genomics of the *Campylobacter lari* group." *Genome Biol Evol* 6 (12):3252-66. doi: 10.1093/gbe/evu249.

Minami, M., M. Ohta, T. Ohkura, T. Ando, K. Torii, T. Hasegawa, and H. Goto. 2006. "Use of a combination of brushing technique and the loop-mediated isothermal amplification method as a novel, rapid, and safe system for detection of *Helicobacter pylori*." *J Clin Microbiol* 44 (11):4032-7. doi: 10.1128/jcm.00898-06.

Miroshnichenko, M. L., N. A. Kostrikina, S. L'Haridon, C. Jeanthon, H. Hippe, E. Stackebrandt, and E. A. Bonch-Osmolovskaya. 2002. "*Nautilia lithotrophica* gen. nov., sp. nov., a thermophilic sulfur-reducing epsilon-proteobacterium isolated from a deep-sea hydrothermal vent." *Int J Syst Evol Microbiol* 52 (Pt 4):1299-304. doi: 10.1099/ijs.0.02139-0.

Miroshnichenko, M. L., S. L'Haridon, P. Schumann, S. Spring, E. A. Bonch-Osmolovskaya, C. Jeanthon, and E. Stackebrandt. 2004. "*Caminibacter profundus* sp. nov., a novel

thermophile of *Nautiliales* ord. nov. within the class '*Epsilonproteobacteria*', isolated from a deep-sea hydrothermal vent." *Int J Syst Evol Microbiol* 54 (Pt 1):41-5. doi: 10.1099/ijs.0.02753-0.

Miyagawa, J., H. Maeda, T. Murauchi, S. Kokeguchi, K. Yamabe, I. Tanimoto, F. Nishimura, K. Fukui, and S. Takashiba. 2008. "Rapid and simple detection of eight major periodontal pathogens by the loop-mediated isothermal amplification method." *FEMS Immunol Med Microbiol* 53 (3):314-21. doi: 10.1111/j.1574-695X.2008.00417.x.

Moore, E. R. B., S. A. Mihaylova, P. Vandamme, M. I. Krichevsky, and L. Dijkshoorn. 2010. "Microbial systematics and taxonomy: relevance for a microbial commons." *Res Microbiol* 161 (6):430-8. doi: 10.1016/j.resmic.2010.05.007.

Moorthie, S., C. J. Mattocks, and C. F. Wright. 2011. "Review of massively parallel DNA sequencing technologies." *Hugo J* 5 (1-4):1-12. doi: 10.1007/s11568-011-9156-3.

Mower, W. R. 1999. "Evaluating bias and variability in diagnostic test reports." *Ann Emerg Med* 33 (1):85-91.

MRC-Holland. 2014. "Designing synthetic MLPA probes." In. Amsterdam, The Netherlands: MRC-Holland (accessed 31/1/2015).

MRC-Holland. 2015. "Designing synthetic MLPA probes." In. Amsterdam, The Netherlands: MRC-Holland (accessed 18/3/2016).

Mukherjee, S., M. Huntemann, N. Ivanova, N. C. Kyrpides, and A. Pati. 2015. "Large-scale contamination of microbial isolate genomes by Illumina PhiX control." *Stand Genomic Sci* 10:18. doi: 10.1186/1944-3277-10-18.

Murray, P. R. 2010. "Matrix-assisted laser desorption ionization time-of-flight mass spectrometry: usefulness for taxonomy and epidemiology." *Clin Microbiol Infect* 16 (11):1626-30. doi: 10.1111/j.1469-0691.2010.03364.x.

Murray, R. G. E., and K. H. Schleifer. 1994. "Taxonomic notes: a proposal for recording the properties of putative taxa of procaryotes." *Int J Syst Bacteriol* 44 (1):174-6. doi: 10.1099/00207713-44-1-174.

Musher, D. M., and B. L. Musher. 2004. "Contagious acute gastrointestinal infections." *N Engl J Med* 351 (23):2417-27. doi: 10.1056/NEJMra041837.

Musmanno, R. A., M. Russi, N. Figura, P. Guglielmetti, A. Zanchi, R. Signori, and A. Rossolini. 1998. "Unusual species of campylobacters isolated in the Siena Tuscany area, Italy." *New Microbiol* 21 (1):15-22.

Nagarajan, N., and M. Pop. 2013. "Sequence assembly demystified." *Nat Rev Genet* 14 (3):157-67. doi: 10.1038/nrg3367.

Nakagawa, S., F. Inagaki, K. Takai, K. Horikoshi, and Y. Sako. 2005. "*Thioreductor micantisoli* gen. nov., sp. nov., a novel mesophilic, sulfur-reducing chemolithoautotroph within the *epsilon-Proteobacteria* isolated from hydrothermal sediments in the Mid-Okinawa Trough." *Int J Syst Evol Microbiol* 55 (Pt 2):599-605. doi: 10.1099/ijs.0.63351-0.

Nakagawa, S., and K. Takai. 2014. "The family *Nautiliaceae*: The genera *Caminibacter, Lebetimonas*, and *Nautilia*." In *The Prokaryotes: Deltaproteobacteria and*

Epsilonproteobacteria, edited by Eugene Rosenberg, Edward F. DeLong, Stephen Lory, Erko Stackebrandt and Fabiano Thompson, 393-399. Berlin, Heidelberg: Springer Berlin Heidelberg.

Nakagawa, S., K. Takai, F. Inagaki, K. Horikoshi, and Y. Sako. 2005. "*Nitratiruptor tergarcus* gen. nov., sp. nov. and *Nitratifractor salsuginis* gen. nov., sp. nov., nitrate-reducing chemolithoautotrophs of the *epsilon-Proteobacteria* isolated from a deep-sea hydrothermal system in the Mid-Okinawa Trough." *Int J Syst Evol Microbiol* 55 (Pt 2):925-33. doi: 10.1099/ijs.0.63480-0.

Nielsen, H. L., T. Ejlertsen, J. Engberg, and H. Nielsen. 2013. "High incidence of *Campylobacter concisus* in gastroenteritis in North Jutland, Denmark: a population-based study." *Clin Microbiol Infect* 19 (5):445-50. doi: 10.1111/j.1469-0691.2012.03852.x.

Nielsen, H. L., J. Engberg, T. Ejlertsen, and H. Nielsen. 2013a. "Clinical manifestations of *Campylobacter concisus* infection in children." *Pediatr Infect Dis J* 32 (11):1194-8. doi: 10.1097/INF.0b013e31829f0aff.

Nielsen, H. L., J. Engberg, T. Ejlertsen, and H. Nielsen. 2013b. "Comparison of polycarbonate and cellulose acetate membrane filters for isolation of *Campylobacter concisus* from stool samples." *Diagn Microbiol Infect Dis* 76 (4):549-50. doi: 10.1016/j.diagmicrobio.2013.05.002.

Nielsen, H. L., H. Nielsen, and M. Torpdahl. 2016. "Multilocus sequence typing of *Campylobacter concisus* from Danish diarrheic patients." *Gut Pathog* 8:44. doi: 10.1186/s13099-016-0126-0.

O'Leary, N. A., M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt. 2016. "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation." *Nucleic Acids Res* 44 (D1):D733-45. doi: 10.1093/nar/gkv1189.

Olesen, B., J. Neimann, B. Bottiger, S. Ethelberg, P. Schiellerup, C. Jensen, M. Helms, F. Scheutz, K. E. Olsen, K. Krogfelt, E. Petersen, K. Molbak, and P. Gerner-Smidt. 2005. "Etiology of diarrhea in young children in Denmark: a case-control study." *J Clin Microbiol* 43 (8):3636-41. doi: 10.1128/jcm.43.8.3636-3641.2005.

On, S. L. W. 1994. "Confirmation of human *Campylobacter concisus* isolates misidentified as *Campylobacter mucosalis* and suggestions for improved differentiation between the two species." *J Clin Microbiol* 32 (9):2305-6.

On, S. L. W. 1996. "Identification methods for campylobacters, helicobacters, and related organisms." *Clin Microbiol Rev* 9 (3):405-22.

On, S. L. W. 2001. "Taxonomy of *Campylobacter, Arcobacter*, *Helicobacter* and related bacteria: current status, future prospects and immediate concerns." *J Appl Microbiol* 90 (S6):1S-15S. doi: 10.1046/j.1365-2672.2001.01349.x.

On, S. L. W. 2005. "Taxonomy, phylogeny, and methods for the identification of *Campylobacter* species." In *Campylobacter: Molecular and Cellular Biology*, edited by J. M. Ketley and M. E. Kondel, 13-42. Wymondham, UK: Horizon Bioscience.

On, S. L. W. 2013. "Isolation, identification and subtyping of *Campylobacter*: where to from here?" *J Microbiol Methods* 95 (1):3-7. doi: 10.1016/j.mimet.2013.06.011.

On, S. L. W., S. M. Brandt, A. J. Cornelius, V. Fusco, G. M. Quero, E. Mackiw, K. Houf, A. Bilbao, A. I. Diaz, L. Benejat, F. Megraud, J. Collins-Emerson, N. P. French, V. Gotcheva, A. Angelov, H.-L. Alakomi, M. Saarela, and S. M. Paulin. 2013. "PCR revisited: a case for revalidation of PCR assays for microorganisms using identification of *Campylobacter* species as an exemplar." *Qual Assur Saf Crops Foods* 5 (1):49-62. doi: 10.3920/QAS2012.0158.

On, S. L. W., and A. J. Cornelius. 2016. "*Campylobacter*: properties and occurence." In *Encyclopedia of Food and Health*, edited by B. Caballero, P. Finglas and F. Toldra, 602-8. Oxford: Academic Press.

On, S. L. W., and B. Holmes. 1995. "Classification and identification of campylobacters, helicobacters and allied taxa by numerical analysis of phenotypic characters." *Syst Appl Microbiol* 18:374-390. doi: 10.1016/S0723-2020(11)80431-5.

On, S. L. W., A. Lee, J. L O'Rourke, F. E. Dewhirst, B. J. Paster, J. G. Fox, and P. Vandamme. 2005. "Genus I. *Helicobacter*." In *Bergey's Manual of Systematic Bacteriology*, edited by D. J. Brenner, N. R. Krieg and J. T. Staley, 1169-89. USA: Springer.

On, S. L. W., W. G. Miller, K. Houf, J. G. Fox, and P. Vandamme. 2017. "Minimal standards for describing new species belonging to the families *Campylobacteraceae* and *Helicobacteraceae*: *Campylobacter*, *Arcobacter*, *Helicobacter* and *Wolinella* spp." *Int J Syst Evol Microbiol*. doi: 10.1099/ijsem.0.002255.

On, S. L. W., B. L. Siemer, S. M. Brandt, P. Chung, and A. J. Lastovica. 2013. "Characterisation of *Campylobacter concisus* strains from South Africa using amplified fragment length polymorphism (AFLP) profiling and a genomospecies-specific polymerase chain reaction (PCR) assay: identification of novel genomospecies and correlation with clinical data." *Afr J Microbiol Res* 7:1845-1851. doi: 10.5897/AJMR12.2182.

Oren, A., and G. M. Garrity. 2017. "List of novel names and novel combinations previously effectively, but not validly, published." *Int J Syst Evol Microbiol* 67 (7):2075-2078. doi: doi:10.1099/ijsem.0.002122.

Oren, Aharon, and George M. Garrity. 2015. "List of new names and new combinations previously effectively, but not validly, published." *Int J Syst Evol Microbiol* 65 (7):2017-2025. doi: doi:10.1099/ijs.0.000317.

Overbeek, R., T. Begley, R. M. Butler, J. V. Choudhuri, H. Y. Chuang, M. Cohoon, V. de Crecy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G. D. Pusch, D. A. Rodionov, C. Ruckert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, and V. Vonstein. 2005. "The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes." *Nucleic Acids Res* 33 (17):5691-702. doi: 10.1093/nar/gki866.

Owen, R. J. 1998. "*Helicobacter* - species classification and identification." *Br Med Bull* 54 (1):17-30. doi: 10.1093/oxfordjournals.bmb.a011667.

Oyama, K., S. Khan, T. Okamoto, S. Fujii, K. Ono, T. Matsunaga, J. Yoshitake, T. Sawa, J. Tomida, Y. Kawamura, and T. Akaike. 2012. "Identification of and screening for human *Helicobacter cinaedi* infections and carriers via nested PCR." *J Clin Microbiol* 50 (12):3893-900. doi: 10.1128/jcm.01622-12.

Ozer, E. A., J. P. Allen, and A. R. Hauser. 2014. "Characterization of the core and accessory genomes of *Pseudomonas aeruginosa* using bioinformatic tools Spine and AGEnt." *BMC Genomics* 15:737. doi: 10.1186/1471-2164-15-737.

Page, A. J., C. A. Cummins, M. Hunt, V. K. Wong, S. Reuter, M. T. G. Holden, M. Fookes, D. Falush, J. A. Keane, and J. Parkhill. 2015. "Roary: rapid large-scale prokaryote pan genome analysis." *Bioinformatics* 31 (22):3691-3. doi: 10.1093/bioinformatics/btv421.

Pantoja, Y., K. Pinheiro, A. Veras, F. Araujo, A. Lopes de Sousa, L. C. Guimaraes, A. Silva, and R. T. J. Ramos. 2017. "PanWeb: A web interface for pan-genomic analysis." *PLoS One* 12 (5):e0178154. doi: 10.1371/journal.pone.0178154.

Parker, C. T., W. G. Miller, S. T. Horn, and A. J. Lastovica. 2007. "Common genomic features of *Campylobacter jejuni* subsp. *doylei* strains distinguish them from *C. jejuni* subsp. *jejuni*." *BMC Microbiol* 7:50. doi: 10.1186/1471-2180-7-50.

Parker, C. T., B. Quinones, W. G. Miller, S. T. Horn, and R. E. Mandrell. 2006. "Comparative genomic analysis of *Campylobacter jejuni* strains reveals diversity due to genomic elements similar to those present in *C. jejuni* strain RM1221." *J Clin Microbiol* 44 (11):4125-35. doi: 10.1128/jcm.01231-06.

Parkhill, J., B. W. Wren, K. Mungall, J. M. Ketley, C. Churcher, D. Basham, T. Chillingworth, R. M. Davies, T. Feltwell, S. Holroyd, K. Jagels, A. V. Karlyshev, S. Moule, M. J. Pallen, C. W. Penn, M. A. Quail, M. A. Rajandream, K. M. Rutherford, A. H. M. van Vliet, S. Whitehead, and B. G. Barrell. 2000. "The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences." *Nature* 403 (6770):665-8. doi: 10.1038/35001088.

Parte, Aidan C. 2014. "LPSN—list of prokaryotic names with standing in nomenclature." *Nucleic Acids Res* 42 (D1):D613-D616. doi: 10.1093/nar/gkt1111.

Paul, S., A. Bhardwaj, S. K. Bag, E. V. Sokurenko, and S. Chattopadhyay. 2015. "PanCoreGen - profiling, detecting, annotating protein-coding genes in microbial genomes." *Genomics* 106 (6):367-72. doi: 10.1016/j.ygeno.2015.10.001.

Pei, A. Y., W. E. Oberdorf, C. W. Nossa, A. Agarwal, P. Chokshi, E. A. Gerz, Z. Jin, P. Lee, L. Yang, M. Poles, S. M. Brown, S. Sotero, T. DeSantis, E. Brodie, K. Nelson, and Z. Pei. 2010. "Diversity of 16S rRNA genes within individual prokaryotic genomes." *Appl Environ Microbiol* 76 (12):3886-97. doi: 10.1128/aem.02953-09.

Perez-Rodriguez, I., J. Ricci, J. W. Voordeckers, V. Starovoytov, and C. Vetriani. 2010. "*Nautilia nitratireducens* sp. nov., a thermophilic, anaerobic, chemosynthetic, nitrate-ammonifying bacterium isolated from a deep-sea hydrothermal vent." *Int J Syst Evol Microbiol* 60 (Pt 5):1182-6. doi: 10.1099/ijs.0.013904-0.

Petersen, R. F., C. S. Harrington, H. E. Kortegaard, and S. L. W. On. 2007. "A PCR-DGGE method for detection and identification of *Campylobacter*, *Helicobacter*, *Arcobacter* and related *Epsilobacteria* and its application to saliva samples from humans and domestic pets." *J Appl Microbiol* 103 (6):2601-15. doi: 10.1111/j.1365-2672.2007.03515.x.

Pham, N. T. K., Q. D. Trinh, P. Khamrin, N. Ukarapol, T. Kongsricharoern, W. Yamazaki, S. Komine-Aizawa, S. Okitsu, N. Maneekarn, S. Hayakawa, and H. Ushijima. 2015. "Loop-mediated isothermal amplification (LAMP) for detection of *Campylobacter jejuni* and *C. coli* in Thai children with diarrhea." *Jpn J Infect Dis* 68 (5):432-3. doi: 10.7883/yoken.JJID.2014.450.

Piccirillo, A., G. Niero, L. Calleros, R. Perez, H. Naya, and G. Iraola. 2016. "*Campylobacter geochelonis* sp. nov. isolated from the western Hermann's tortoise (*Testudo hermanni hermanni*)." *Int J Syst Evol Microbiol* 66 (9):3468-74. doi: 10.1099/ijsem.0.001219.

Pride, D. T., R. J. Meinersmann, T. M. Wassenaar, and M. J. Blaser. 2003. "Evolutionary implications of microbial genome tetranucleotide frequency biases." *Genome Res* 13 (2):145-58. doi: 10.1101/gr.335003.

R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rasko, D. A., G. S. A. Myers, and J. Ravel. 2005. "Visualization of comparative genomic analyses by BLAST score ratio." *BMC Bioinformatics* 6:2. doi: 10.1186/1471-2105-6-2.

Read, D. S., D. J. Woodcock, N. J. Strachan, K. J. Forbes, F. M. Colles, M. C. J. Maiden, F. Clifton-Hadley, A. Ridley, A. Vidal, J. Rodgers, A. S. Whiteley, and S. K. Sheppard. 2013. "Evidence for phenotypic plasticity among multihost *Campylobacter jejuni* and *C. coli* lineages, obtained using ribosomal multilocus sequence typing and Raman spectroscopy." *Appl Environ Microbiol* 79 (3):965-73. doi: 10.1128/aem.02521-12.

Revez, J., M. Rossi, S. Piva, D. Florio, A. Lucchi, A. Parisi, G. Manfreda, and R. G. Zanoni. 2013. "Occurrence of epsilon-proteobacterial species in rabbits (*Oryctolagus cuniculus*)

reared in intensive and rural farms." *Vet Microbiol* 162 (1):288-92. doi: 10.1016/j.vetmic.2012.08.009.

Revez, J., T. Schott, M. Rossi, and M. L. Hanninen. 2012. "Complete genome sequence of a variant of *Campylobacter jejuni* NCTC 11168." *J Bacteriol* 194 (22):6298-9. doi: 10.1128/jb.01385-12.

Richter, M., and R. Rossello-Mora. 2009. "Shifting the genomic gold standard for the prokaryotic species definition." *Proc Natl Acad Sci U S A* 106 (45):19126-31. doi: 10.1073/pnas.0906412106.

Ricker, N., H. Qian, and R. R. Fulthorpe. 2012. "The limitations of draft assemblies for understanding prokaryotic adaptation and evolution." *Genomics* 100 (3):167-75. doi: 10.1016/j.ygeno.2012.06.009.

Rimbara, E., M. Sasatsu, and D. Y. Graham. 2013. "PCR detection of *Helicobacter pylori* in clinical samples." *Methods Mol Biol* 943:279-87. doi: 10.1007/978-1-60327-353-4_19.

Rivers, T. M. 1937. "Viruses and Koch's postulates." *J Bacteriol* 33 (1):1-12.

Robertson, L. A., J. G. Kuenen, B. J. Paster, F. E. Dewhirst, and P. Vandamme. 2005. "Genus II. *Thiovulum*." In *Bergey's Manual of Systematic Bacteriology*, edited by D. J. Brenner, N. R. Krieg and J. T. Staley, 1189-91. USA: Springer.

Rossello-Mora, R., and R. Amann. 2015. "Past and future species definitions for *Bacteria* and *Archaea*." *Syst Appl Microbiol* 38 (4):209-16. doi: 10.1016/j.syapm.2015.02.001.

Rossen, L., P. Norskov, K. Holmstrom, and O. F. Rasmussen. 1992. "Inhibition of PCR by components of food samples, microbial diagnostic assays and DNA-extraction solutions." *Int J Food Microbiol* 17 (1):37-45. doi: 10.1016/0168-1605(92)90017-W.

Rossi, M., L. Debruyne, R. G. Zanoni, G. Manfreda, J. Revez, and P. Vandamme. 2009. "*Campylobacter avium* sp. nov., a hippurate-positive species isolated from poultry." *Int J Syst Evol Microbiol* 59 (Pt 9):2364-9. doi: 10.1099/ijs.0.007419-0.

Sahl, J. W., C. J. Allender, R. E. Colman, K. J. Califf, J. M. Schupp, B. J. Currie, K. E. Van Zandt, H. C. Gelhaus, P. Keim, and A. Tuanyok. 2015. "Genomic characterization of *Burkholderia pseudomallei* isolates selected for medical countermeasures testing: comparative genomics associated with differential virulence." *PLoS One* 10 (3):e0121052. doi: 10.1371/journal.pone.0121052.

Sahl, J. W., J. G. Caporaso, D. A. Rasko, and P. Keim. 2014. "The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes." *PeerJ* 2:e332. doi: 10.7717/peerj.332.

Sahl, J. W., J. D. Gillece, J. M. Schupp, V. G. Waddell, E. M. Driebe, D. M. Engelthaler, and P. Keim. 2013. "Evolution of a pathogen: a comparative genomics analysis identifies a genetic pathway to pathogenesis in *Acinetobacter*." *PLoS One* 8 (1):e54287. doi: 10.1371/journal.pone.0054287.

Sahl, J. W., J. R. Sistrunk, N. I. Baby, Y. Begum, Q. Luo, A. Sheikh, F. Qadri, J. M. Fleckenstein, and D. A. Rasko. 2017. "Insights into enterotoxigenic *Escherichia coli*

diversity in Bangladesh utilizing genomic epidemiology." *Sci Rep* 7 (1):3402. doi: 10.1038/s41598-017-03631-x.

Salama, S. M., H. Tabor, M. Richter, and D. E. Taylor. 1992. "Pulsed-field gel electrophoresis for epidemiologic studies of *Campylobacter hyointestinalis* isolates." *J Clin Microbiol* 30 (8):1982-4.

Salipante, S. J., D. J. Roach, J. O. Kitzman, M. W. Snyder, B. Stackhouse, S. M. Butler-Wu, C. Lee, B. T. Cookson, and J. Shendure. 2015. "Large-scale genomic sequencing of extraintestinal pathogenic *Escherichia coli* strains." *Genome Res* 25 (1):119-28. doi: 10.1101/gr.180190.114.

Salzberg, S. L., A. M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. J. Treangen, M. C. Schatz, A. L. Delcher, M. Roberts, G. Marcais, M. Pop, and J. A. Yorke. 2012. "GAGE: A critical evaluation of genome assemblies and assembly algorithms." *Genome Res* 22 (3):557-67. doi: 10.1101/gr.131383.111.

Samie, A., C. L. Obi, L. J. Barrett, S. M. Powell, and R. L. Guerrant. 2007. "Prevalence of *Campylobacter* species, *Helicobacter pylori* and *Arcobacter* species in stool samples from the Venda region, Limpopo, South Africa: studies using molecular diagnostic methods." *J Infect* 54 (6):558-66. doi: 10.1016/j.jinf.2006.10.047.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. "DNA sequencing with chain-terminating inhibitors." *Proc Natl Acad Sci U S A* 74 (12):5463-7.

Sansom, C. 2000. "Database searching with DNA and protein sequences: an introduction." *Brief Bioinform* 1 (1):22-32. doi: 10.1093/bib/1.1.22.

Sasi Jyothsna, T. S., K. Rahul, E. V. V. Ramaprasad, Ch. Sasikala, and Ch. V. Ramana. 2013. "*Arcobacter anaerophilus* sp. nov., isolated from an estuarine sediment and emended description of the genus *Arcobacter*." *Int J Syst Evol Microbiol* 63 (Pt 12):4619-25. doi: 10.1099/ijs.0.054155-0.

Scanlon, K. A., C. Cagney, D. Walsh, D. McNulty, A. Carroll, E. B. McNamara, D. A. McDowell, and G. Duffy. 2013. "Occurrence and characteristics of fastidious *Campylobacteraceae* species in porcine samples." *Int J Food Microbiol* 163 (1):6-13. doi: 10.1016/j.ijfoodmicro.2013.02.004.

Schouten, J. P., C. J. McElgunn, R. Waaijer, D. Zwijnenburg, F. Diepvens, and G. Pals. 2002. "Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification." *Nucleic Acids Res* 30 (12):e57. doi: 10.1093/nar/gnf056.

Scott, D., and B. Ely. 2015. "Comparison of genome sequencing technology and assembly methods for the analysis of a GC-rich bacterial genome." *Curr Microbiol* 70 (3):338-44. doi: 10.1007/s00284-014-0721-6.

Seemann, T. 2014. "Prokka: rapid prokaryotic genome annotation." *Bioinformatics* 30 (14):2068-9. doi: 10.1093/bioinformatics/btu153.

Serraino, A., D. Florio, F. Giacometti, S. Piva, D. Mion, and R. G. Zanoni. 2013. "Presence of *Campylobacter* and *Arcobacter* species in in-line milk filters of farms authorized to

produce and sell raw milk and of a water buffalo dairy farm in Italy." *J Dairy Sci* 96 (5):2801-7. doi: 10.3168/jds.2012-6249.

Shao, Y., X. He, E. M. Harrison, C. Tai, H. Y. Ou, K. Rajakumar, and Z. Deng. 2010. "mGenomeSubtractor: a web-based tool for parallel *in silico* subtractive hybridization analysis of multiple bacterial genomes." *Nucleic Acids Res* 38 (Web Server issue):W194-200. doi: 10.1093/nar/gkq326.

Shea, S., K. A. Kubota, H. Maguire, S. Gladbach, A. Woron, R. Atkinson-Dunn, M. R. Couturier, and M. B. Miller. 2017. "Clinical microbiology laboratories' adoption of culture-independent diagnostic tests is a threat to foodborne-disease surveillance in the United States." *J Clin Microbiol* 55 (1):10-19. doi: 10.1128/jcm.01624-16.

Shen, Z., Y. Feng, S. Muthupalani, A. Sheh, L. E. Cheaney, C. A. Kaufman, G. Gong, B. J. Paster, and J. G. Fox. 2016. "Novel *Helicobacter* species *H. japonicum* isolated from laboratory mice from Japan induces typhlocolitis and lower bowel carcinoma in C57BL/129 IL10-/- mice." *Carcinogenesis* 37 (12):1190-1198. doi: 10.1093/carcin/bgw101.

Shen, Z., Y. Feng, A. Sheh, J. Everitt, F. Bertram, B. J. Paster, and J. G. Fox. 2015. "Isolation and characterization of a novel *Helicobacter* species, *Helicobacter jaachi* sp. nov., from common marmosets (*Callithrix jaachus*)." *J Med Microbiol* 64 (9):1063-73. doi: 10.1099/jmm.0.000113.

Shen, Z., A. Mannion, M. T. Whary, S. Muthupalani, A. Sheh, Y. Feng, G. Gong, P. Vandamme, H. R. Holcombe, B. J. Paster, and J. G. Fox. 2016. "*Helicobacter saguini*, a novel *Helicobacter* isolated from cotton-top tamarins with ulcerative colitis, has proinflammatory properties and induces typhlocolitis and dysplasia in gnotobiotic IL-10-/- mice." *Infect Immun* 84 (8):2307-16. doi: 10.1128/iai.00235-16.

Sheppard, S. K., X. Didelot, G. Meric, A Torralba, K. A. Jolley, D. J. Kelly, S. D. Bentley, M. C. J. Maiden, J. Parkhill, and D. Falush. 2013. Genome-wide association study identifies vitamin B5 synthesis as a host specificity factor in *Campylobacter*. Dryad Digital Repository.

Sheppard, S. K., X. Didelot, G. Meric, A. Torralbo, K. A. Jolley, D. J. Kelly, S. D. Bentley, M. C. J. Maiden, J. Parkhill, and D. Falush. 2013. "Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*." *Proc Natl Acad Sci U S A* 110 (29):11923-7. doi: 10.1073/pnas.1305559110.

Silman, Alan J, and Gary J Macfarlane. 2002. *Epidemiological Studies: a Practical Guide*. Second ed. Cambridge, United Kingdom: Cambridge University Press.

Simpson, J. T., and M. Pop. 2015. "The theory and practice of genome sequence assembly." *Annu Rev Genomics Hum Genet* 16:153-72. doi: 10.1146/annurev-genom-090314-050032.

Sims, G. E., S. R. Jun, G. A. Wu, and S. H. Kim. 2009. "Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions." *Proc Natl Acad Sci U S A* 106 (8):2677-82. doi: 10.1073/pnas.0813249106.

Skarp-de Haan, C. P. A., A. Culebro, T. Schott, J. Revez, E. K. H. Schweda, M. L. Hanninen, and M. Rossi. 2014. "Comparative genomics of unintrogressed *Campylobacter coli* clades 2 and 3." *BMC Genomics* 15:129. doi: 10.1186/1471-2164-15-129.

Smith, J. L., B. J. Campbell, T. E. Hanson, C. L. Zhang, and S. C. Cary. 2008. "*Nautilia profundicola* sp. nov., a thermophilic, sulfur-reducing epsilonproteobacterium from deep-sea hydrothermal vents." *Int J Syst Evol Microbiol* 58 (Pt 7):1598-602. doi: 10.1099/ijs.0.65435-0.

Sorokin, D. Y., T. P. Tourova, and G. Muyzer. 2013. "Isolation and characterization of two novel alkalitolerant sulfidogens from a Thiopaq bioreactor, *Desulfonatronum alkalitolerans* sp. nov., and *Sulfurospirillum alkalitolerans* sp. nov." *Extremophiles* 17 (3):535-43. doi: 10.1007/s00792-013-0538-4.

Stackebrandt, E., W. Frederiksen, G. M. Garrity, P. A. D. Grimont, P. Kampfer, M. C. J. Maiden, X. Nesme, R. Rossello-Mora, J. Swings, H. G. Truper, L. Vauterin, A. C. Ward, and W. B. Whitman. 2002. "Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology." *Int J Syst Evol Microbiol* 52 (Pt 3):1043-7. doi: 10.1099/ijs.0.02360-0.

Stackebrandt, E., and B. M. Goebel. 1994. "Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology." *Int J Syst Bacteriol* 44 (4):846-849. doi: 10.1099/00207713-44-4-846.

Steinbrueckner, B., G. Haerter, K. Pelz, S. Weiner, J. A. Rump, W. Deissler, S. Bereswill, and M. Kist. 1997. "Isolation of *Helicobacter pullorum* from patients with enteritis." *Scand J Infect Dis* 29 (3):315-8. doi: 10.3109/00365549709019053.

Stewart, A. C., B. Osborne, and T. D. Read. 2009. "DIYA: a bacterial annotation pipeline for any genomics lab." *Bioinformatics* 25 (7):962-3. doi: 10.1093/bioinformatics/btp097.

Stolz, J. F., R. S. Oremland, B. J. Paster, F. E. Dewhirst, and P. Vandamme. 2005. "Genus III. *Sulfurospirillum*." In *Bergey's Manual of Systematic Bacteriology*, edited by D. J. Brenner, N. R. Krieg and J. T. Staley, 1165-8. USA: Springer.

Stothard, P., and D. S. Wishart. 2006. "Automated bacterial genome analysis and annotation." *Curr Opin Microbiol* 9 (5):505-10. doi: 10.1016/j.mib.2006.08.002.

Suerbaum, S., G. Geis, C. Josenhans, and W. Opferkuch. 1992. "Biochemical studies of *Helicobacter mustelae* fatty acid composition and flagella." *Infect Immun* 60 (4):1695-8.

Taboada, E. N., M. R. Graham, J. A. Carrico, and G. Van Domselaar. 2017. "Food safety in the age of next generation sequencing, bioinformatics, and open data access." *Front Microbiol* 8:909. doi: 10.3389/fmicb.2017.00909.

Takai, K., H. Hirayama, T. Nakagawa, Y. Suzuki, K. H. Nealson, and K. Horikoshi. 2005. "*Lebetimonas acidiphila* gen. nov., sp. nov., a novel thermophilic, acidophilic, hydrogen-oxidizing chemolithoautotroph within the '*Epsilonproteobacteria*', isolated from a deep-sea hydrothermal fumarole in the Mariana Arc." *Int J Syst Evol Microbiol* 55 (Pt 1):183-9. doi: 10.1099/ijs.0.63330-0.

Takai, K., K. H. Nealson, and K. Horikoshi. 2004. "*Hydrogenimonas thermophila* gen. nov., sp. nov., a novel thermophilic, hydrogen-oxidizing chemolithoautotroph within the *epsilon-Proteobacteria*, isolated from a black smoker in a Central Indian Ridge hydrothermal field." *Int J Syst Evol Microbiol* 54 (Pt 1):25-32. doi: 10.1099/ijs.0.02787-0.

Takai, K., M. Suzuki, S. Nakagawa, M. Miyazaki, Y. Suzuki, F. Inagaki, and K. Horikoshi. 2006. "*Sulfurimonas paralvinellae* sp. nov., a novel mesophilic, hydrogen- and sulfur-oxidizing chemolithoautotroph within the *Epsilonproteobacteria* isolated from a deep-sea hydrothermal vent polychaete nest, reclassification of *Thiomicrospira denitrificans* as *Sulfurimonas denitrificans* comb. nov. and emended description of the genus *Sulfurimonas*." *Int J Syst Evol Microbiol* 56 (Pt 8):1725-33. doi: 10.1099/ijs.0.64255-0.

Tam, C. C., S. J. O'Brien, D. S. Tompkins, F. J. Bolton, L. Berry, J. Dodds, D. Choudhury, F. Halstead, M. Iturriza-Gomara, K. Mather, G. Rait, A. Ridge, L. C. Rodrigues, J. Wain, B. Wood, J. J. Gray, and the IID2 Study Executive Committee. 2012. "Changes in causes of acute gastroenteritis in the United Kingdom over 15 years: microbiologic findings from 2 prospective, population-based studies of infectious intestinal disease." *Clin Infect Dis* 54 (9):1275-86. doi: 10.1093/cid/cis028.

Taniguchi, T., A. Sekiya, M. Higa, Y. Saeki, K. Umeki, A. Okayama, T. Hayashi, and N. Misawa. 2014. "Rapid identification and subtyping of *Helicobacter cinaedi* strains by intact-cell mass spectrometry profiling with the use of matrix-assisted laser desorption ionization-time of flight mass spectrometry." *J Clin Microbiol* 52 (1):95-102. doi: 10.1128/jcm.01798-13.

Tanner, A. C. R., S. Badger, C. H. Lai, M. A. Listgarten, R. A. Visconti, and S. S. Socransky. 1981. "*Wolinella* gen. nov., *Wolinella succinogenes* (*Vibrio succinogenes* Wolin et al.) comb. nov., and description of *Bacteriodes gracilis* sp. nov., *Wolinella recta* sp. nov., *Campylobacter concisus* sp. nov., and *Eikenella corrodens* from humans with periodontal disease." *Int J Syst Bacteriol* 31 (4):432-445. doi: 10.1099/00207713-31-4-432.

Tatusov, R. L., E. V. Koonin, and D. J. Lipman. 1997. "A genomic perspective on protein families." *Science* 278 (5338):631-7. doi: 10.1126/science.278.5338.631.

Tatusov, R. L., D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova, and E. V. Koonin. 2001. "The COG database: new developments in phylogenetic classification of proteins from complete genomes." *Nucleic Acids Res* 29 (1):22-8. doi: 10.1093/nar/29.1.22.

Tatusova, T., M. DiCuccio, A. Badretdin, V. Chetvernin, E. P. Nawrocki, L. Zaslavsky, A. Lomsadze, K. D. Pruitt, M. Borodovsky, and J. Ostell. 2016. "NCBI prokaryotic genome annotation pipeline." *Nucleic Acids Res* 44 (14):6614-24. doi: 10.1093/nar/gkw569.

Taylor, D. N., J. A. Kiehlbauch, W. Tee, C. Pitarangsi, and P. Echeverria. 1991. "Isolation of group 2 aerotolerant *Campylobacter* species from Thai children with diarrhea." *J Infect Dis* 163 (5):1062-7. doi: 10.1093/infdis/163.5.1062.

Taylor, M. R., J. A. Conrad, D. Wahl, and P. J. O'Brien. 2011. "Kinetic mechanism of human DNA ligase I reveals magnesium-dependent changes in the rate-limiting step that compromise ligation efficiency." *J Biol Chem* 286 (26):23054-62. doi: 10.1074/jbc.M111.248831.

Tazumi, A., S. Nakanishi, S. Meguro, Y. Kakinuma, J. E. Moore, B. C. Millar, and M. Matsuda. 2010. "Occurrence and characterisation of intervening sequences (IVSs) within 16S rRNA genes from two atypical *Campylobacter* species, *C. sputorum and C. curvus*." *Br J Biomed Sci* 67 (2):77-81. doi: 10.1080/09674845.2010.11730295.

Teeling, H., A. Meyerdierks, M. Bauer, R. Amann, and F. O. Glockner. 2004. "Application of tetranucleotide frequencies for the assignment of genomic fragments." *Environ Microbiol* 6 (9):938-47. doi: 10.1111/j.1462-2920.2004.00624.x.

Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. Deboy, T. M. Davidsen, M. Mora, M. Scarselli, I. Margarit y Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. B. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli, and C. M. Fraser. 2005. "Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome"." *Proc Natl Acad Sci U S A* 102 (39):13950-5. doi: 10.1073/pnas.0506758102.

The UniProt Consortium. 2017. "UniProt: the universal protein knowledgebase." *Nucleic Acids Res* 45 (D1):D158-d169. doi: 10.1093/nar/gkw1099.

Thepault, A., G. Meric, K. Rivoal, B. Pascoe, L. Mageiros, F. Touzain, V. Rose, V. Beven, M. Chemaly, and S. K. Sheppard. 2017. "Genome-wide identification of host-segregating epidemiological markers for source attribution in *Campylobacter jejuni*." *Appl Environ Microbiol* 83 (7):e03085-16. doi: 10.1128/aem.03085-16.

Thompson, C. C., L. Chimetto, R. A. Edwards, J. Swings, E. Stackebrandt, and F. L. Thompson. 2013. "Microbial genomic taxonomy." *BMC Genomics* 14:913. doi: 10.1186/1471-2164-14-913.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." *Nucleic Acids Res* 22 (22):4673-80. doi: 10.1093/nar/22.22.4673.

Tian, R. M., L. Cai, W. P. Zhang, H. L. Cao, and P. Y. Qian. 2015. "Rare events of intragenus and intraspecies horizontal transfer of the 16S rRNA gene." *Genome Biol Evol* 7 (8):2310-20. doi: 10.1093/gbe/evv143.

Tilmanne, A., D. Martiny, H. Hallin, A. Cornelius, M. Wautier, C. Quash, P. Lepage, and O. Vandenberg. 2018. "*Campylobacter concisus* and acute gastroenteritis in children: lack

of association." *Pediatr Infect Dis J* Published online ahead of print. doi: 10.1097/inf.0000000000002028.

Tindall, B. J., R. Rossello-Mora, H. J. Busse, W. Ludwig, and P. Kampfer. 2010. "Notes on the characterization of prokaryote strains for taxonomic purposes." *Int J Syst Evol Microbiol* 60 (Pt 1):249-66. doi: 10.1099/ijs.0.016949-0.

Tompkins, D. S., M. J. Hudson, H. R. Smith, R. P. Eglin, J. G. Wheeler, M. M. Brett, R. J. Owen, J. S. Brazier, P. Cumberland, V. King, and P. E. Cook. 1999. "A study of infectious intestinal disease in England: microbiological findings in cases and controls." *Commun Dis Public Health* 2 (2):108-13.

Underwood, A. P., N. O. Kaakoush, N. Sodhi, J. Merif, W. Seah Lee, S. M. Riordan, W. D. Rawlinson, and H. M. Mitchell. 2016. "*Campylobacter concisus* pathotypes are present at significant levels in patients with gastroenteritis." *J Med Microbiol* 65 (3):219-26. doi: 10.1099/jmm.0.000216.

Ursing, J. B., H. Lior, and R. J. Owen. 1994. "Proposal of minimal standards for describing new species of the family *Campylobacteraceae*." *Int J Syst Bacteriol* 44 (4):842-5. doi: 10.1099/00207713-44-4-842.

Ushijima, H., S. Nishimura, A. Thongprachum, Y. Shimizu-Onda, D. N. Tran, N. T. K. Pham, S. Takanashi, S. K. Dey, S. Okitsu, W. Yamazaki, M. Mizuguchi, and S. Hayakawa. 2014. "Sensitive and rapid detection of *Campylobacter* species from stools of children with diarrhea in Japan by the loop-mediated isothermal amplification method." *Jpn J Infect Dis* 67 (5):374-8. doi: 10.7883/yoken.67.374.

Vallenet, D., S. Engelen, D. Mornico, S. Cruveiller, L. Fleury, A. Lajus, Z. Rouy, D. Roche, G. Salvignol, C. Scarpelli, and C. Medigue. 2009. "MicroScope: a platform for microbial genome annotation and comparative genomics." *Database (Oxford)* 2009:bap021. doi: 10.1093/database/bap021.

Vallenet, D., L. Labarre, Z. Rouy, V. Barbe, S. Bocs, S. Cruveiller, A. Lajus, G. Pascal, C. Scarpelli, and C. Medigue. 2006. "MaGe: a microbial genome annotation system supported by synteny results." *Nucleic Acids Res* 34 (1):53-65. doi: 10.1093/nar/gkj406.

Vallenet, David, Alexandra Calteau, Stéphane Cruveiller, Mathieu Gachet, Aurélie Lajus, Adrien Josso, Jonathan Mercier, Alexandre Renaux, Johan Rollin, Zoe Rouy, David Roche, Claude Scarpelli, and Claudine Médigue. 2017. "MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes." *Nucleic Acids Research* 45 (D1):D517-D528. doi: 10.1093/nar/gkw1101.

Van den Abeele, A. M., D. Vogelaers, J. Van Hende, and K. Houf. 2014. "Prevalence of *Arcobacter* species among humans, Belgium, 2008-2013." *Emerg Infect Dis* 20 (10):1731-4. doi: 10.3201/eid2010.140433.

Van Etterijck, R., J. Breynaert, H. Revets, T. Devreker, Y. Vandenplas, P. Vandamme, and S. Lauwers. 1996. "Isolation of *Campylobacter concisus* from feces of children with and without diarrhea." *J Clin Microbiol* 34 (9):2304-6.

Van, T. T. H., E. Elshagmani, M. C. Gor, P. C. Scott, and R. J. Moore. 2016. "*Campylobacter hepaticus* sp. nov., isolated from chickens with spotty liver disease." *Int J Syst Evol Microbiol* 66 (11):4518-4524. doi: 10.1099/ijsem.0.001383.

van Vliet, A. H. M., and J. G. Kusters. 2015. "Use of alignment-free phylogenetics for rapid genome sequence-based typing of *Helicobacter pylori* virulence markers and antibiotic susceptibility." *J Clin Microbiol* 53 (9):2877-88. doi: 10.1128/jcm.01357-15.

Vandamme, P., L. Debruyne, E. De Brandt, and E. Falsen. 2010. "Reclassification of *Bacteroides ureolyticus* as *Campylobacter ureolyticus* comb. nov., and emended description of the genus *Campylobacter*." *Int J Syst Evol Microbiol* 60 (Pt 9):2016-22. doi: 10.1099/ijs.0.017152-0.

Vandamme, P., F. E. Dewhirst, B. J. Paster, and S. L. W. On. 2005a. "Family I. *Campylobacteraceae*." In *Bergey's Manual of Systematic Bacteriology*, edited by D. J. Brenner, N. R. Krieg and J. T. Staley, 1145-6. USA: Springer.

Vandamme, P., F. E. Dewhirst, B. J. Paster, and S. L. W. On. 2005b. "Genus I. *Campylobacter*." In *Bergey's Manual of Systematic Bacteriology*, edited by D. J. Brenner, N. R. Krieg and J. T. Staley, 1147-60. USA: Springer.

Vandamme, P., F. E. Dewhirst, B. J. Paster, and S. L. W. On. 2005c. "Genus II. *Arcobacter*." In *Bergey's Manual of Systematic Bacteriology*, edited by D. J. Brenner, N. R. Krieg and J. T. Staley, 1161-5. USA: Springer.

Vandamme, P., E. Falsen, B. Pot, B. Hoste, K. Kersters, and J. De Ley. 1989. "Identification of EF group 22 campylobacters from gastroenteritis cases as *Campylobacter concisus*." *J Clin Microbiol* 27 (8):1775-81.

Vandamme, P., C. S. Harrington, K. Jalava, and S. L. W. On. 2000. "Misidentifying helicobacters: the *Helicobacter cinaedi* example." *J Clin Microbiol* 38 (6):2261-6.

Vandamme, P., and C. Peeters. 2014. "Time to revisit polyphasic taxonomy." *Antonie Van Leeuwenhoek* 106 (1):57-65. doi: 10.1007/s10482-014-0148-x.

Vandamme, P., B. Pot, E. Falsen, K. Kersters, and J. de Ley. 1990. "Intra- and interspecific relationships of veterinary campylobacters revealed by numerical analysis of electrophoretic profiles and DNA:DNA hybridizations." *Syst Appl Microbiol* 13:295-303. doi: 10.1016/S0723-2020(11)80201-8.

Vandamme, P., B. Pot, M. Gillis, P. de Vos, K. Kersters, and J. Swings. 1996. "Polyphasic taxonomy, a consensus approach to bacterial systematics." *Microbiol Rev* 60 (2):407-38.

Vandenberg, O., A. J. Cornelius, H. Souayah, D. Martiny, L. Vlaes, S. M. Brandt, and S. L. W. On. 2013. "The role of Epsilonproteobacteria in children with gastroenteritis." *Pediatr Infect Dis J* 32 (10):1140-2. doi: 10.1097/INF.0b013e3182975047.

Vandenberg, O., A. Dediste, K. Houf, S. Ibekwem, H. Souayah, S. Cadranel, N. Douat, G. Zissis, J. P. Butzler, and P. Vandamme. 2004. "*Arcobacter* species in humans." *Emerg Infect Dis* 10 (10):1863-7. doi: 10.3201/eid1010.040241.

Varghese, N. J., S. Mukherjee, N. Ivanova, K. T. Konstantinidis, K. Mavrommatis, N. C. Kyrpides, and A. Pati. 2015. "Microbial species delineation using whole genome sequences." *Nucleic Acids Res* 43 (14):6761-71. doi: 10.1093/nar/gkv657.

Voordeckers, J. W., V. Starovoytov, and C. Vetriani. 2005. "*Caminibacter mediatlanticus* sp. nov., a thermophilic, chemolithoautotrophic, nitrate-ammonifying bacterium isolated from a deep-sea hydrothermal vent on the Mid-Atlantic Ridge." *Int J Syst Evol Microbiol* 55 (Pt 2):773-9. doi: 10.1099/ijs.0.63430-0.

Waite, D. W., I. Vanwonterghem, C. Rinke, D. H. Parks, Y. Zhang, K. Takai, S. M. Sievert, J. Simon, B. J. Campbell, T. E. Hanson, T. Woyke, M. G. Klotz, and P. Hugenholtz. 2017. "Comparative genomic analysis of the class *Epsilonproteobacteria* and proposed reclassification to Epsilonbacteraeota (phyl. nov.)." *Front Microbiol* 8:682. doi: 10.3389/fmicb.2017.00682.

Wallace, L., S. C. Daugherty, S. Nagaraj, J. K. Johnson, A. D. Harris, and D. A. Rasko. 2016. "Use of comparative genomics to characterize the diversity of *Acinetobacter baumannii* surveillance isolates in a health care institution." *Antimicrob Agents Chemother* 60 (10):5933-41. doi: 10.1128/aac.00477-16.

Wang, G., C. G. Clark, T. M. Taylor, C. Pucknell, C. Barton, L. Price, D. L. Woodward, and F. G. Rodgers. 2002. "Colony multiplex PCR assay for identification and differentiation of *Campylobacter jejuni*, *C. coli*, *C. lari*, *C. upsaliensis*, and *C. fetus* subsp. *fetus*." *J Clin Microbiol* 40 (12):4744-7. doi: 10.1128/jcm.40.12.4744–4747.2002.

Wang, Y., F. Liu, X. Zhang, H. K. L. Chung, S. M. Riordan, M. C. Grimm, S. Zhang, R. Ma, S. A. Lee, and L. Zhang. 2017. "*Campylobacter concisus* genomospecies 2 is better adapted to the human gastrointestinal tract as compared with *Campylobacter concisus* genomospecies 1." *Front Physiol* 8:543. doi: 10.3389/fphys.2017.00543.

Wareing, D. R. A., S. T. Aspinall, P. G. Hayward, and D. N. Hutchinson. 1998. "Improved selective medium (CAT) for thermophilic campylobacters including *Campylobacter upsaliensis*." In *Campylobacter, Helicobacter and Related Organisms. Proceedings of the 9th International Workshop*, edited by A. J. Lastovica, D. G. Newell and E. E. Lastovica, 46-49. Cape Town, South Africa: Institute of Child Health.

Wattam, A. R., D. Abraham, O. Dalay, T. L. Disz, T. Driscoll, J. L. Gabbard, J. J. Gillespie, R. Gough, D. Hix, R. Kenyon, D. Machi, C. Mao, E. K. Nordberg, R. Olson, R. Overbeek, G. D. Pusch, M. Shukla, J. Schulman, R. L. Stevens, D. E. Sullivan, V. Vonstein, A. Warren, R. Will, M. J. C. Wilson, H. S. Yoo, C. Zhang, Y. Zhang, and B. W. Sobral. 2014. "PATRIC, the bacterial bioinformatics database and analysis resource." *Nucleic Acids Res* 42 (Database issue):D581-91. doi: 10.1093/nar/gkt1099.

Wattam, A. R., J. J. Davis, R. Assaf, S. Boisvert, T. Brettin, C. Bun, N. Conrad, E. M. Dietrich, T. Disz, J. L. Gabbard, S. Gerdes, C. S. Henry, R. W. Kenyon, D. Machi, C. Mao, E. K. Nordberg, G. J. Olsen, D. E. Murphy-Olson, R. Olson, R. Overbeek, B. Parrello, G. D. Pusch, M. Shukla, V. Vonstein, A. Warren, F. Xia, H. Yoo, and R. L. Stevens. 2017.

"Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center." *Nucleic Acids Res* 45 (D1):D535-d542. doi: 10.1093/nar/gkw1017.

Wayne, L. G., D. J. Brenner, R. R. Colwell, P. A. D. Grimont, O. Kandler, M. I. Krichevsky, L. H. Moore, W. E. C. Moore, R. G. E. Murray, E. Stackebrandt, M. P. Starr, and H. G. Truper. 1987. "Report of the ad hoc committee of reconciliation of approaches to bacterial systematics." *Int J Syst Bacteriol* 37 (4):463-464. doi: 10.1099/00207713-37-4-463.

Webb, A. L., V. F. Boras, P. Kruczkiewicz, L. B. Selinger, E. N. Taboada, and G. D. Inglis. 2016. "Comparative detection and quantification of *Arcobacter butzleri* in stools from diarrheic and nondiarrheic people in Southwestern Alberta, Canada." *J Clin Microbiol* 54 (4):1082-8. doi: 10.1128/jcm.03202-15.

Wheeler, J. G., D. Sethi, J. M. Cowden, P. G. Wall, L. C. Rodrigues, D. S. Tompkins, M. J. Hudson, and P. J. Roderick. 1999. "Study of infectious intestinal disease in England: rates in the community, presenting to general practice, and reported to national surveillance." *BMJ* 318 (7190):1046-50. doi: 10.1136/bmj.318.7190.1046.

Wheeler, N. E., L. Barquist, R. A. Kingsley, and P. P. Gardner. 2016. "A profile-based method for identifying functional divergence of orthologous genes in bacterial genomes." *Bioinformatics* 32 (23):3566-3574. doi: 10.1093/bioinformatics/btw518.

Whiteduck-Leveillee, K., J. Whiteduck-Leveillee, M. Cloutier, J. T. Tambong, R. Xu, E. Topp, M. T. Arts, J. Chao, Z. Adam, C. Andre Levesque, D. R. Lapen, R. Villemur, G. Talbot, and I. U. H. Khan. 2015. "*Arcobacter lanthieri* sp. nov., isolated from pig and dairy cattle manure." *Int J Syst Evol Microbiol* 65 (8):2709-16. doi: 10.1099/ijs.0.000318.

Wilson, A. C., S. S. Carlson, and T. J. White. 1977. "Biochemical evolution." *Annu Rev Biochem* 46:573-639. doi: 10.1146/annurev.bi.46.070177.003041.

Winkler, M. A., J. Uher, and S. Cepa. 1999. "Direct analysis and identification of *Helicobacter* and *Campylobacter* species by MALDI-TOF mass spectrometry." *Anal Chem* 71 (16):3416-9. doi: 10.1021/ac990135r.

Woese, C. R. 1987. "Bacterial evolution." *Microbiol Rev* 51 (2):221-71.

Wong, T. L., M. L. Devane, J. A. Hudson, P. Scholes, M. G. Savill, and J. D. Klena. 2004. "Validation of a PCR method for *Campylobacter* detection on poultry packs." *Br Food J* 106 (9):642-650. doi: 10.1108/00070700410558175.

Wybo, I., J. Breynaert, S. Lauwers, F. Lindenburg, and K. Houf. 2004. "Isolation of *Arcobacter skirrowii* from a patient with chronic diarrhea." *J Clin Microbiol* 42 (4):1851-2. doi: 10.1128/jcm.42.4.1851-1852.2004.

Yamazaki-Matsune, W., M. Taguchi, K. Seto, R. Kawahara, K. Kawatsu, Y. Kumeda, M. Kitazato, M. Nukina, N. Misawa, and T. Tsukamoto. 2007. "Development of a multiplex PCR assay for identification of *Campylobacter coli*, *Campylobacter fetus*, *Campylobacter hyointestinalis* subsp. *hyointestinalis*, *Campylobacter jejuni, Campylobacter lari* and *Campylobacter upsaliensis*." *J Med Microbiol* 56 (Pt 11):1467-73. doi: 10.1099/jmm.0.47363-0.

Yamazaki, W., M. Taguchi, M. Ishibashi, M. Kitazato, M. Nukina, N. Misawa, and K. Inoue. 2008. "Development and evaluation of a loop-mediated isothermal amplification assay for rapid and simple detection of *Campylobacter jejuni* and *Campylobacter coli*." *J Med Microbiol* 57 (Pt 4):444-51. doi: 10.1099/jmm.0.47688-0.

Yari, F., R. Abiri, E. Aryan, T. Ahmadi Jouybari, J. Navabi, and A. Alvandi. 2016. "Loop-mediated isothermal amplification as a fast noninvasive method of *Helicobacter pylori* diagnosis." *J Clin Lab Anal* 30 (5):464-70. doi: 10.1002/jcla.21880.

Yoshizaki, S., T. Umemura, K. Tanaka, K. Watanabe, M. Hayashi, and Y. Muto. 2014. "Genome-wide evidence of positive selection in *Bacteroides fragilis*." *Comput Biol Chem* 52:43-50. doi: 10.1016/j.compbiolchem.2014.09.001.

You, Y., C. Fu, X. Zeng, D. Fang, X. Yan, B. Sun, D. Xiao, and J. Zhang. 2008. "A novel DNA microarray for rapid diagnosis of enteropathogenic bacteria in stool specimens of patients with diarrhea." *J Microbiol Methods* 75 (3):566-71. doi: 10.1016/j.mimet.2008.09.007.

Yuvaraj, I., J. Sridhar, D. Michael, and K. Sekar. 2017. "PanGeT: Pan-genomics tool." *Gene* 600:77-84. doi: 10.1016/j.gene.2016.11.025.

Zanoni, R. G., L. Debruyne, M. Rossi, J. Revez, and P. Vandamme. 2009. "*Campylobacter cuniculorum* sp. nov., from rabbits." *Int J Syst Evol Microbiol* 59 (Pt 7):1666-71. doi: 10.1099/ijs.0.007286-0.

Zerbino, D. R., and E. Birney. 2008. "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." *Genome Res* 18 (5):821-9. doi: 10.1101/gr.074492.107.

Zhang, H., S. Morrison, and Y. W. Tang. 2015. "Multiplex polymerase chain reaction tests for detection of pathogens associated with gastroenteritis." *Clin Lab Med* 35 (2):461-86. doi: 10.1016/j.cll.2015.02.006.

Zhang, L., H. Lee, M. C. Grimm, S. M. Riordan, A. S. Day, and D. A. Lemberg. 2014. "*Campylobacter concisus* and inflammatory bowel disease." *World J Gastroenterol* 20 (5):1259-67. doi: 10.3748/wjg.v20.i5.1259.

Zhang, L., S. M. Man, A. S. Day, S. T. Leach, D. A. Lemberg, S. Dutt, M. Stormon, A. Otley, E. V. O'Loughlin, A. Magoffin, P. H. Y. Ng, and H. Mitchell. 2009. "Detection and isolation of *Campylobacter* species other than *C. jejuni* from children with Crohn's disease." *J Clin Microbiol* 47 (2):453-5. doi: 10.1128/jcm.01949-08.

Zhang, Z., C. Yu, X. Wang, S. Yu, and X. H. Zhang. 2016. "*Arcobacter pacificus* sp. nov., isolated from seawater of the South Pacific Gyre." *Int J Syst Evol Microbiol* 66:542-7. doi: 10.1099/ijsem.0.000751.

Zhao, Y., J. Wu, J. Yang, S. Sun, J. Xiao, and J. Yu. 2012. "PGAP: pan-genomes analysis pipeline." *Bioinformatics* 28 (3):416-8. doi: 10.1093/bioinformatics/btr655.

Zuker, M. 2003. "Mfold web server for nucleic acid folding and hybridization prediction." *Nucleic Acids Res* 31 (13):3406-15. doi: 10.1093/nar/gkg595.

Supplementary Information

Supplementary File 1

Functional Group	COG Code	Ccon_10_1_50	Ccon_13826	Ccon_ATCC33237	Ccon_ATCC51561	Ccon_ATCC51562	Ccon_AUS22_Bd2	Ccon_CCUG19995	Ccon_Lasto127_99	Ccon_Lasto205_94	Ccon_Lasto220_96	Ccon_Lasto28_99	Ccon_Lasto393_96	Ccon_Lasto61_99	Ccon_Lasto64_99	Ccon_RCH26	Ccon_RMIT_JF1	Ccon_RMIT_O17	Ccon_UNSW1	Ccon_UNSW2	Ccon_UNSW3	Ccon_UNSWCD	Ccon_UNSWCS	Ccur_DSM6644	Cfet_82_40	Chom_ATCC_BAA_38	Cjej_ATCC33560	Cmuc_CCUG21559	Cmuc_DSM21682	Crec_ATCC33238	Csho_ATCC51146	Cspu_CCUG20703
RNA processing and modification	A	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Energy production and conversion	C	131	130	119	140	123	124	137	134	120	119	121	122	118	124	147	125	127	136	130	128	124	131	156	123	104	115	113	116	119	122	142
Cell cycle control, cell division, chromosome partitioning	D	22	21	20	22	20	20	22	23	24	20	20	23	21	22	23	24	20	23	21	21	20	24	18	15	20	15	19	19	24	22	22
Amino acid transport and metabolism	E	113	119	109	108	104	105	122	112	107	106	103	104	104	104	118	108	105	111	114	106	106	112	112	121	93	121	120	130	113	111	108
Nucleotide transport and metabolism	F	49	49	45	49	45	44	53	48	49	46	50	46	44	45	50	46	46	48	50	47	47	47	47	45	45	48	49	55	49	50	49
Carbohydrate transport and metabolism	G	44	46	43	46	43	44	46	50	41	45	44	45	43	46	49	46	41	48	47	48	44	46	51	46	32	55	50	49	45	41	37
Coenzyme transport and metabolism	H	80	77	75	78	80	81	80	82	81	80	78	81	79	78	86	81	81	81	83	81	83	82	75	86	76	78	80	81	91	80	79
Lipid transport and metabolism	I	41	45	42	43	43	41	48	43	42	43	41	41	45	42	46	46	44	47	44	40	41	43	55	40	38	36	41	43	59	48	39
Translation, ribosomal structure and biogenesis	J	172	179	170	168	172	167	171	170	164	167	166	166	168	164	182	174	173	165	177	168	166	179	165	168	161	164	175	186	179	162	164
Transcription	K	49	46	39	45	35	38	42	46	43	31	43	43	38	42	42	46	39	44	47	47	40	61	45	42	33	35	48	45	63	52	37
Replication, recombination and repair	L	84	81	88	76	84	83	91	85	94	80	99	77	80	87	102	93	82	84	95	95	77	105	73	71	91	76	92	78	97	87	72
Cell wall/membrane/envelope biogenesis	M	134	136	129	130	128	124	141	129	131	130	138	130	130	128	137	127	129	130	125	127	124	127	134	124	120	126	131	141	133	127	120
Cell motility	N	36	37	36	38	38	35	36	35	35	36	36	34	35	35	42	36	35	39	37	35	35	36	38	35	5	36	39	44	35	34	38
Posttranslational modification, protein turnover, chaperones	O	98	101	95	97	90	93	98	103	87	88	92	92	92	92	101	97	99	94	99	97	92	100	94	83	75	77	95	92	108	89	94
Inorganic ion transport and metabolism	P	86	85	101	78	103	105	81	81	101	98	101	100	98	106	111	103	98	88	87	86	83	80	108	103	56	79	79	88	110	109	84
Secondary metabolites biosynthesis, transport and catabolism	Q	14	21	12	17	9	7	16	21	12	8	13	8	11	13	13	14	9	19	17	15	11	17	19	14	20	13	14	13	31	26	6
General function prediction only	R	113	127	109	120	110	109	118	123	105	113	104	110	109	112	122	118	115	110	114	108	97	123	141	111	97	106	120	127	135	120	102
Function unknown	S	81	87	73	86	79	78	87	87	82	78	84	72	82	81	88	89	77	82	86	88	84	95	82	63	59	75	78	75	102	96	67
Signal transduction mechanisms	T	66	69	61	63	63	57	60	58	62	55	63	58	59	60	72	58	58	54	59	59	58	66	61	37	37	40	59	61	92	83	55
Intracellular trafficking, secretion, and vesicular transport	U	24	46	38	38	31	35	38	33	34	28	38	38	35	37	43	36	36	37	33	26	25	30	23	29	40	40	32	30	52	27	36
Defense mechanisms	V	49	44	34	46	35	43	42	43	41	46	43	39	42	37	44	39	35	47	51	42	34	60	44	38	35	39	36	34	62	44	31
Mobilome: prophages, transposons	X	13	13	25	9	24	5	9	13	35	2	63	24	14	6	19	34	11	6	16	35	14	52	8	8	26	4	37	16	38	7	9
Cytoskeleton	Z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1
no definition	BQ	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0
no definition	CE	1	1	1	1	1	1	1	1	2	1	1	1	0	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
no definition	CO	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no definition	CP	7	7	7	7	7	7	8	5	7	7	7	7	7	7	6	7	7	5	6	7	7	7	7	8	5	5	6	7	5	7	9
no definition	DL	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
no definition	DM	2	2	2	2	2	2	2	3	2	2	3	2	2	2	2	2	2	2	2	2	3	2	3	2	2	2	2	4	3	3	2
no definition	DN	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
no definition	DP	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0	1	1
no definition	DZ	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0
no definition	EF	3	3	2	3	2	2	3	3	2	2	2	2	2	2	3	3	2	1	3	4	3	2	2	3	2	2	3	4	2	2	3
no definition	EG	2	2	1	2	1	1	3	2	1	1	1	1	1	1	1	1	1	2	2	2	2	2	1	1	1	1	1	1	1	1	1
no definition	EH	6	6	6	6	6	6	6	6	7	6	7	6	6	6	6	7	6	6	6	8	7	6	9	9	6	7	8	9	7	6	8
no definition	EM	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1
no definition	EP	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	0	4	4	5	0	2	7
no definition	EQ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
no definition	ER	3	3	4	3	2	2	2	3	2	2	2	2	2	2	4	2	2	3	3	3	3	3	2	2	1	1	2	2	3	3	2
no definition	ET	8	6	6	6	6	7	6	6	5	6	6	5	5	6	7	7	7	6	6	4	7	6	4	4	7	4	3	3	3	6	1
no definition	FE	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
no definition	FR	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	2	2	2
no definition	FV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
no definition	GI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
no definition	GM	2	5	3	2	4	3	3	3	3	3	3	3	3	3	3	3	3	3	2	3	3	4	4	3	2	3	3	4	4	3	
no definition	GR	1	1	1	1	1	1	1	1	1	1	2	1	1	2	2	1	1	1	1	1	1	1	1	0	0	0	2	2	1	1	1

Supplementary File 1

Functional Group	COG Code	Ccon_10_1_50	Ccon_13826	Ccon_ATCC33237	Ccon_ATCC51561	Ccon_ATCC51562	Ccon_AUS22_Bd2	Ccon_CCUG19995	Ccon_Lasto127_99	Ccon_Lasto205_94	Ccon_Lasto220_96	Ccon_Lasto28_99	Ccon_Lasto393_96	Ccon_Lasto61_99	Ccon_Lasto64_99	Ccon_RCH26	Ccon_RMIT_JF1	Ccon_RMIT_O17	Ccon_UNSW1	Ccon_UNSW2	Ccon_UNSW3	Ccon_UNSWCD	Ccon_UNSWCS	Ccur_DSM6644	Cfet_82_40	Chom_ATCC_BAA_38:	Cjej_ATCC33560	Cmuc_CCUG21559	Cmuc_DSM21682	Crec_ATCC33238	Csho_ATCC51146	Cspu_CCUG20703
no definition	GT	1	1	0	1	0	0	1	1	0	0	0	0	1	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0
no definition	HC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
no definition	HE	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
no definition	HI	3	3	3	3	3	3	3	3	3	3	3	3	3	4	3	4	3	3	3	3	3	3	3	5	3	3	3	4	3	3
no definition	HJ	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	1	2	1
no definition	HQ	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no definition	HR	5	5	3	4	3	3	4	5	3	4	3	3	3	3	5	3	3	4	5	4	5	4	3	4	3	5	3	3	4	3	3
no definition	HT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
no definition	IQ	3	9	9	10	10	3	9	10	4	9	4	4	10	9	9	10	9	9	6	4	3	4	11	4	2	5	11	11	10	10	3
no definition	IR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0
no definition	JM	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	2	0	1	0	0	0
no definition	JO	1	0	1	1	2	2	3	1	1	1	1	2	2	2	2	2	2	1	1	1	1	1	0	1	0	1	1	1	1	1	1
no definition	JR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	1	1	1
no definition	JT	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1
no definition	JU	1	3	1	1	1	1	2	2	1	1	1	1	1	1	1	1	1	3	1	1	1	1	1	1	1	0	1	1	4	1	4
no definition	KE	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
no definition	KG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
no definition	KJ	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1
no definition	KL	0	0	2	1	0	0	0	0	0	1	0	1	0	0	0	2	1	1	0	0	0	0	1	0	3	0	1	0	3	0	0
no definition	KN	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1	1	0	0	1	1	1	1	1	1	0	1	1	1	1	1	1
no definition	KR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	5	1	1	2	1	1
no definition	KT	1	2	1	2	1	2	2	1	4	3	6	2	1	1	2	3	1	3	1	1	1	2	2	1	3	2	3	1	4	3	2
no definition	KX	0	1	0	1	1	1	1	1	3	1	3	0	0	0	2	1	0	0	2	1	1	1	0	0	0	0	1	0	1	4	0
no definition	LK	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
no definition	LX	4	9	4	5	3	3	5	4	7	2	11	4	4	6	9	6	4	3	6	3	2	3	3	1	4	3	5	3	6	4	1
no definition	M-	1	1	1	1	1	1	1	1	1	2	1	2	2	2	1	1	2	1	1	1	3	1	1	1	1	0	1	1	1	1	1
no definition	MI	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	9	1	1	2	2	1	1	2	
no definition	MN	2	2	2	2	2	2	2	2	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
no definition	MO	2	3	2	4	3	2	4	3	2	3	3	3	3	2	3	2	3	4	3	2	2	2	2	2	2	2	3	4	2	2	2
no definition	MR	2	3	2	2	2	3	2	2	2	3	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	4	2	2	2
no definition	MU	3	2	2	1	4	2	1	1	2	3	2	2	2	2	2	2	2	1	1	2	2	1	1	3	3	5	2	2	2	1	2
no definition	MV	7	7	6	7	7	6	7	7	6	8	8	8	7	6	7	6	6	7	8	7	7	8	7	6	4	4	5	5	7	8	7
no definition	MX	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
no definition	NT	21	18	14	17	14	14	18	23	16	15	14	14	14	15	14	13	12	14	16	14	13	17	14	25	2	18	21	19	15	11	24
no definition	NU	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	8	7	7	7	7	7	7	1	8	7	7	7	7	8
no definition	NW	7	7	6	6	7	5	7	7	5	5	5	5	5	5	6	5	5	6	8	6	6	8	7	6	5	1	4	5	7	5	8
no definition	OK	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
no definition	OT	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0
no definition	PC	1	2	0	1	0	0	2	2	1	0	1	0	0	0	0	0	0	1	2	1	1	1	2	1	1	0	1	1	0	1	1
no definition	PH	2	2	2	1	4	4	1	1	2	2	2	3	5	5	3	2	3	2	1	1	1	1	6	3	0	1	2	2	6	7	2
no definition	PR	1	0	1	1	0	0	2	1	1	0	1	1	0	0	0	0	1	0	0	0	0	1	1	1	1	0	1	0	1	2	0
no definition	PT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
no definition	PV	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
no definition	QC	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
no definition	QR	1	2	1	1	2	1	1	1	1	1	1	1	2	2	2	1	1	1	1	1	1	2	2	4	1	5	1	2	2	3	1
no definition	QT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Supplementary File 1

Functional Group	COG Code	Ccon_10_1_50	Ccon_13826	Ccon_ATCC33237	Ccon_ATCC51561	Ccon_ATCC51562	Ccon_AUS22_Bd2	Ccon_CCUG19995	Ccon_Lasto127_99	Ccon_Lasto205_94	Ccon_Lasto220_96	Ccon_Lasto28_99	Ccon_Lasto393_96	Ccon_Lasto61_99	Ccon_Lasto64_99	Ccon_RCH26	Ccon_RMIT_JF1	Ccon_RMIT_O17	Ccon_UNSW1	Ccon_UNSW2	Ccon_UNSW3	Ccon_UNSWCD	Ccon_UNSWCS	Ccur_DSM6644	Cfet_82_40	Chom_ATCC_BAA_38:	Cjej_ATCC33560	Cmuc_CCUG21559	Cmuc_DSM21682	Crec_ATCC33238	Csho_ATCC51146	Cspu_CCUG20703
no definition	QV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0
no definition	TE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
no definition	TK	18	22	19	21	19	21	21	24	19	21	21	19	20	19	20	20	17	22	25	23	22	23	28	21	11	12	23	23	27	25	23
no definition	UW	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	3	0	2	2	0	0	0
no definition	CHR	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
no definition	CIR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
no definition	DPM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0
no definition	FGR	4	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
no definition	FTP	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	2
no definition	GER	4	4	4	4	4	4	4	4	4	4	5	4	4	4	4	4	4	4	5	4	4	4	6	4	4	2	7	7	5	6	4
no definition	IQR	3	4	4	3	4	2	4	3	3	3	2	2	4	3	4	3	4	4	5	2	2	2	5	5	3	3	4	4	3	3	3
no definition	MDT	2	2	1	3	1	1	1	3	1	1	1	2	2	2	1	2	1	1	3	3	1	3	2	1	0	1	1	1	2	2	1
no definition	NUW	5	6	5	6	5	5	6	6	6	5	6	5	5	5	5	6	5	6	7	5	5	6	5	13	11	4	5	4	7	7	6
no definition	UXR	0	2	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	1	0	0
no definition	GEPR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	2	1	1	1	1	1	1	2	0	9	1	1	1	1	1
no definition	n/d*	273	366	247	288	245	242	405	324	276	204	329	243	248	260	299	341	240	255	350	297	231	419	218	168	284	223	344	230	485	334	221

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	content	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
A. aquimarinus	RM 15229^T	RM15229	Dr William Miller	1	Complete	2520808	26.8	2489	139	15	56	1	EpsiloFsa, Epsilo Faa	Arco (ignored)
A. bivalviorum	RM 15224^T	RM15224	Dr William Miller	1	Complete	2684689	28.11	2622	189	12	47	1	EpsiloFsa, Epsilo Faa	Arco
A. butzleri	7h1h	CP006615	GenBank	1	Complete	2253233	27.06	2207	132	15	55	1	EpsiloFsa, Epsilo Faa	Arco
A. butzleri	ED-1	AP012047	GenBank	1	Complete	2256675	27.07	2167	125	15	55	1	EpsiloFsa, Epsilo Faa	Arco
A. butzleri	JV22	AEPT01000000	GenBank	78	Draft	2253392	26.84	2316	118	3	45	1	EpsiloFsa, Epsilo Faa	Arco
A. butzleri	RM4018^T	CP000361	GenBank	1	Complete	2341251	27.05	2279	141	15	55	1	EpsiloFsa, Epsilo Faa	Arco
A. cibarius	LMG 21996^T	JABW01000000	GenBank	44	Draft	2201349	27.12	2168	109	14	47	1	EpsiloFsa, Epsilo Faa	Arco
A. cibarius	RM 15232	RM15232	Dr William Miller	1	Complete	2343493	27.15	2342	123	18	53	1	EpsiloFsa, Epsilo Faa	Arco
A. cloacae	RM 15227^T	RM15227	Dr William Miller	1	Complete	2621880	27	2613	154	15	60	1	EpsiloFsa, Epsilo Faa	Arco
A. cryaerophilus	RM 1582^T	RM1582	Dr William Miller	1	Complete	2002416	27.45	1992	134	15	49	1	EpsiloFsa, Epsilo Faa	Arco
A. cryaerophilus	RM 5557	RM5557	Dr William Miller	1	Complete	2055914	27.51	2019	127	15	51	1	EpsiloFsa, Epsilo Faa	Arco
A. defluvii	RM 14018^T	RM14018	Dr William Miller	1	Complete	3002244	26.42	2980	181	16	53	1	EpsiloFsa, Epsilo Faa	Arco
A. ellisii	RM 15222^T	RM15222	Dr William Miller	1	Complete	2787287	26.88	2822	169	15	60	1	EpsiloFsa, Epsilo Faa	Arco
"A. faecis"	AF 1078	JARS01000000	GenBank	55	Draft	2496785	27.18	2442	147	21	54	1	EpsiloFsa, Epsilo Faa	Arco (ignored)
A. halophilus	RM 5350^T	RM5350	Dr William Miller	1	Complete	2802737	27.6	2671	181	18	55	1	EpsiloFsa, Epsilo Faa	Arco
*A. lanthieri**	AF 1430	JATO01000000	GenBank	28	Draft	2236138	26.41	2226	145	4	40	1	EpsiloFsa, Epsilo Faa	Arco (ignored)
*A. lanthieri**	AF 1440^T	JARU01000000	GenBank	29	Draft	2287768	26.66	2246	150	18	52	1	EpsiloFsa, Epsilo Faa	Arco (ignored)
*A. lanthieri**	AF 1581	JARV01000000	GenBank	24	Draft	2259947	26.84	2218	143	28	57	1	EpsiloFsa, Epsilo Faa	Arco (ignored)
A. molluscorum	RM 14015^T	RM14015	Dr William Miller	1	Complete	2787920	26.21	2722	157	18	55	1	EpsiloFsa, Epsilo Faa	Arco
A. mytili	RM 14013^T	RM14013	Dr William Miller	1	Complete	2849278	26.6	2722	149	18	59	1	EpsiloFsa, Epsilo Faa	Arco
A. nitrofigilis	DSM 7299^T	CP001999	GenBank	1	Complete	3192235	28.36	3146	195	12	56	1	EpsiloFsa, Epsilo Faa	Arco
A. skirrowii	RM 3222^T	RM3222	Dr William Miller	1	Complete	1969755	27.74	1990	93	12	48	1	EpsiloFsa, Epsilo Faa	Arco
A. suis	RM 15228^T	RM15228	Dr William Miller	1	Complete	2624406	27.35	2582	149	15	55	1	EpsiloFsa, Epsilo Faa	Arco
A. thereius	RM 5348^T	RM5348	Dr William Miller	1	Complete	1915080	27	1927	99	12	48	1	EpsiloFsa, Epsilo Faa	Arco
A. trophiarum	RM 12658^T	RM12658	Dr William Miller	1	Complete	1915613	28.22	1927	120	15	48	1	EpsiloFsa, Epsilo Faa	Arco
A. venerupis	RM 16046^T	RM16046	Dr William Miller	1	Complete	3209086	27.36	3132	205	18	64	1	EpsiloFsa, Epsilo Faa	Arco
Arcobacter sp.	AF 1028	JART01000000	GenBank	47	Draft	2414653	27.24	2349	145	18	52	1	EpsiloFsa, Epsilo Faa	Arco (ignored)
Arcobacter sp.	L	AP012048-9	GenBank	2	Complete	2947662	26.58	2874	177	15	56	1	EpsiloFsa, Epsilo Faa	Arco (ignored)
C. avium	RM 8639^T	RM8639	Dr William Miller	1	Complete	1738730	34.17	1737	119	6	40	1	EpsiloFsa, Epsilo Faa	Camp
C. canadensis	RM 9173	RM9173	Dr William Miller	1	Complete	1946154	27.35	1859	116	12	40	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	1091	AIMV01000000	GenBank	132	Draft	1654672	31.33	1659	99	3	30	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	1098	AIMW01000000	GenBank	126	Draft	1740626	31.28	1774	104	3	29	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	1148	AIMX01000000	GenBank	113	Draft	1703486	31.24	1709	102	3	30	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	1417	AIMY01000000	GenBank	97	Draft	1731356	31.32	1796	102	3	33	1	EpsiloFsa, Epsilo Faa	Camp

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	conten	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
C. coli	1891	AINB01000000	GenBank	107	Draft	1497465	31.57	1511	89	3	32	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	1909	AINC01000000	GenBank	133	Draft	1705966	31.36	1718	103	3	30	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	1948	AINE01000000	GenBank	105	Draft	1623842	31.51	1642	95	3	32	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	1957	AINF01000000	GenBank	123	Draft	1735394	31.3	1777	110	3	31	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	1961	AING01000000	GenBank	128	Draft	1733581	31.29	1762	103	3	29	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	2548	AIML01000000	GenBank	162	Draft	1644296	31.31	1659	92	3	32	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	2553	AIMM01000000	GenBank	208	Draft	1825665	31.05	1854	112	3	30	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	2680	AIMN01000000	GenBank	128	Draft	1696901	31.34	1750	99	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	2685	AIMO01000000	GenBank	103	Draft	1713585	31.34	1754	112	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	2688	AIMP01000000	GenBank	209	Draft	1744090	31.26	1788	101	3	28	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	2692	AIMQ01000000	GenBank	168	Draft	1596473	31.47	1605	93	3	32	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	2698	AIMR01000000	GenBank	141	Draft	1515872	31.34	1525	86	3	26	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	76339	HG326877	GenBank	1	Complete	1578522	31.88	1546	103	10	40	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	80352	AIMT01000000	GenBank	260	Draft	1918861	31.05	1946	117	3	32	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	86119	AIMU01000000	GenBank	120	Draft	1736982	31.29	1772	108	3	30	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	111-3	AIMI01000000	GenBank	122	Draft	1709586	31.2	1726	97	3	34	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	132-6	AINA01000000	GenBank	148	Draft	1531994	31.37	1558	100	3	32	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	151-9	AINQ01000000	GenBank	107	Draft	1680317	31.4	1688	100	3	31	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	15-537360	CP006702-3	GenBank	2	Complete	1685020	31.41	1691	110	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	202/04	AINH01000000	GenBank	104	Draft	1675391	31.31	1695	106	3	30	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	317/04	AINJ01000000	GenBank	133	Draft	1753213	31.07	1809	101	3	30	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	37/05	AINK01000000	GenBank	122	Draft	1710249	31.3	1749	106	3	30	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	59-2	AIND01000000	GenBank	140	Draft	1763010	31.12	1793	103	3	31	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	67-8	AINI01000000	GenBank	137	Draft	1502691	31.27	1544	93	3	32	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	7--1	AIMZ01000000	GenBank	99	Draft	1530217	31.34	1543	88	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	84-2	AIMS01000000	GenBank	99	Draft	1605339	31.45	1618	98	3	30	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	90-3	AIMJ01000000	GenBank	121	Draft	1765378	31.22	1798	105	2	30	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	BIGS0001	ANGL01000000	GenBank	167	Draft	1752685	31.43	1746	100	0	37	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	BIGS0002	ANGM01000000	GenBank	181	Draft	1676912	31.38	1690	103	3	35	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	BIGS0003	ANGN01000000	GenBank	118	Draft	1621319	31.81	1573	107	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	BIGS0005	ANGP01000000	GenBank	255	Draft	1589408	31.82	1567	98	3	30	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	BIGS0006	ANGQ01000000	GenBank	65	Draft	1542097	31.86	1510	108	3	32	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	BIGS0007	ANGR01000000	GenBank	197	Draft	1521545	31.99	1471	106	3	30	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	BIGS0008	ANGS01000000	GenBank	478	Draft	3069191	31.95	2964	206	6	64	2	EpsiloFsa, Epsilo Faa	Camp

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	conten	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
C. coli	BIGS0009	ANGT01000000	GenBank	217	Draft	1500025	32.12	1434	103	3	27	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	BIGS0010	ANGU01000000	GenBank	191	Draft	1660889	31.54	1642	103	3	29	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	BIGS0011	ANGV01000000	GenBank	312	Draft	1606337	31.81	1577	97	3	28	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	BIGS0012	ANGW01000000	GenBank	282	Draft	1554713	31.95	1501	91	2	30	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	BIGS0015	ANGZ01000000	GenBank	216	Draft	1594196	31.72	1597	96	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	BIGS0016	ANHA01000000	GenBank	866	Draft	813032	32.47	1160	18	2	14	0	EpsiloFsa, Epsilo Faa	Camp
C. coli	BIGS0017	ANHB01000000	GenBank	79	Draft	1633576	31.49	1661	112	3	35	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	BIGS0018	ANHC01000000	GenBank	206	Draft	1618028	31.59	1679	100	2	30	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	BIGS0019	ANHD01000000	GenBank	276	Draft	1678945	31.55	1694	107	3	27	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	BIGS0020	ANHE01000000	GenBank	691	Draft	1505418	32.02	1490	73	2	19	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	BIGS0021	ANHF01000000	GenBank	374	Draft	1566888	32	1531	87	3	26	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	BIGS0023	ANHH01000000	GenBank	446	Draft	1431051	32.28	1354	84	3	21	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	BIGS0024	ANHI01000000	GenBank	532	Draft	1570976	31.99	1542	81	2	23	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	BIGS0025	ANHJ01000000	GenBank	261	Draft	1657253	31.48	1695	110	3	26	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	BIGS0044	BIGS0044	Dryad	55	Draft	1696727	31.39	1739	107	6	45	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	BIGS0098	BIGS0098	Dryad	119	Draft	1700098	31.51	1702	109	3	35	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	CVM N29710	CP004066-8	GenBank	3	Complete	1732055	31.44	1769	112	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	CVM N29716	ANMS01000000	GenBank	24	Draft	1717288	31.38	1764	115	3	40	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	H56	AINW01000000	GenBank	134	Draft	1736886	31.28	1770	102	3	30	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	H6	AINT01000000	GenBank	122	Draft	1564392	31.48	1604	91	3	29	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	H8	AINU01000000	GenBank	191	Draft	1798181	31	1794	104	3	30	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	H9	AINV01000000	GenBank	91	Draft	1603498	31.34	1617	102	3	29	0	EpsiloFsa, Epsilo Faa	Camp
C. coli	IPSID-1	CBXC01000000	GenBank	39	Draft	1678517	31.36	1714	103	3	41	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	JV20	AEER01000000	GenBank	34	Draft	1678593	31.34	1718	110	3	40	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	K3	AYKN01000000	GenBank	30	Draft	1673801	31.28	1716	106	3	35	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	K7	AYKO01000000	GenBank	51	Draft	1788158	31.2	1847	127	3	43	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	LMG 23336	AINM01000000	GenBank	120	Draft	1648169	31.37	1644	104	3	30	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	LMG 23341	AINN01000000	GenBank	118	Draft	1617023	31.4	1612	101	3	30	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	LMG 23342	AINO01000000	GenBank	109	Draft	1616549	31.39	1620	101	3	30	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	LMG 23344	AINP01000000	GenBank	192	Draft	1784159	31.14	1790	105	3	32	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	LMG 9853	AINR01000000	GenBank	85	Draft	1666318	31.34	1683	108	3	30	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	LMG 9854	AINL01000000	GenBank	118	Draft	1715684	31.25	1728	105	3	30	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	LMG 9860	AINS01000000	GenBank	187	Draft	1851478	31.04	1913	120	3	30	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	RM1875^{T}	CP007183-7	GenBank	5	Complete	1858716	31.06	1954	122	9	44	1	EpsiloFsa, Epsilo Faa	Camp

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	content	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
C. coli	RM2228	AAFL01000000	GenBank	38	Draft	1860666	31.12	1911	121	9	43	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	RM4661	CP007181-2	GenBank	2	Complete	1872235	31.12	1963	112	9	40	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	RM5611	CP007179-80	GenBank	2	Complete	1733165	31.32	1751	107	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	Z156	AINX01000000	GenBank	87	Draft	1686229	31.38	1731	98	3	35	1	EpsiloFsa, Epsilo Faa	Camp
C. coli	Z163	AIMK01000000	GenBank	105	Draft	1677293	31.34	1708	102	3	34	1	EpsiloFsa, Epsilo Faa	Camp
C. concisus GS1	ATCC 33237^T	RM7084	Dr William Miller	1	Complete	1840043	37.62	1848	146	9	47	1	EpsiloFsa, Epsilo Faa	Camp, concisus
C. concisus GS1	ATCC 51562	ANNI01000000	GenBank	21	Draft	1841750	37.69	1841	152	3	39	1	EpsiloFsa, Epsilo Faa	Camp, concisus
C. concisus GS1	AUS22-Bd2	AUS22_Bd2	Dr Mohsina Huq	42	Draft	1835167	37.5	1811	152	3	43	1	RMITConcisus	
C. concisus GS1	Lasto205.94	Lasto205_94	This project	39	Draft	1889689	37.62	1900	146	6	49	1	EpsiloFsa, Epsilo Faa	Camp, concisus
C. concisus GS1	Lasto220.96	Lasto220_96	This project	13	Draft	1795410	37.48	1765	148	6	49	1	EpsiloFsa, Epsilo Faa	Camp, concisus
C. concisus GS1	Lasto28.99	Lasto28_99	This project	30	Draft	1935772	37.73	2016	147	6	49	1	EpsiloFsa, Epsilo Faa	Camp, concisus
C. concisus GS1	Lasto393.96	Lasto393_96	This project	15	Draft	1853764	37.36	1829	149	3	47	1	EpsiloFsa, Epsilo Faa	Camp, concisus
C. concisus GS1	Lasto61.99	Lasto61_99	This project	18	Draft	1835687	37.55	1840	145	6	45	1	EpsiloFsa, Epsilo Faa	Camp, concisus
C. concisus GS1	Lasto64.99	Lasto64_99	This project	20	Draft	1871524	37.49	1852	165	6	49	1	EpsiloFsa, Epsilo Faa	Camp, concisus
C. concisus GS1	RCH 26	RCH26	Dr Mohsina Huq	24	Draft	1915969	37.49	2079	154	3	36	1	RMITConcisus	
C. concisus GS1	RMIT-JF1	RMIT_JF1	Dr Mohsina Huq	25	Draft	1945163	37.62	2003	157	6	44	1	RMITConcisus	
C. concisus GS1	RMIT-O17	RMIT_O17	Dr Mohsina Huq	34	Draft	1838307	37.6	1841	144	3	45	1	RMITConcisus	
C. concisus GS2	13826	CP000792-4	GenBank	3	Complete	2099413	39.27	2072	182	9	47	1	EpsiloFsa, Epsilo Faa	Camp, concisus
C. concisus GS2	ATCC 51561	ANNH01000000	GenBank	69	Draft	1995701	39.64	1936	173	3	41	1	EpsiloFsa, Epsilo Faa	Camp, concisus
C. concisus GS2	CCUG 19995	CCUG19995	This project	44	Draft	2101642	39.39	2091	179	3	44	1	EpsiloFsa, Epsilo Faa	Camp, concisus
C. concisus GS2	Lasto127.99	Lasto127_99	This project	44	Draft	2025154	39.4	1990	155	3	45	1	EpsiloFsa, Epsilo Faa	Camp, concisus
C. concisus GS2	UNSW1	ANNF01000000	GenBank	72	Draft	1937132	39.51	1884	158	3	44	1	EpsiloFsa, Epsilo Faa	Camp, concisus
C. concisus GS2	UNSW2	ANNJ01000000	GenBank	98	Draft	2013661	39.18	2020	147	2	45	1	EpsiloFsa, Epsilo Faa	Camp, concisus
C. concisus GS2	UNSW3	ANNE01000000	GenBank	61	Draft	1907025	39.59	1934	145	3	43	1	EpsiloFsa, Epsilo Faa	Camp, concisus
C. concisus GS2	UNSWCD	AENQ01000000	GenBank	86	Draft	1778912	39.79	1769	138	3	33	1	EpsiloFsa, Epsilo Faa	Camp, concisus
C. concisus GS2	UNSWCS	ANNG01000000	GenBank	177	Draft	2108730	39.25	2165	153	3	42	1	EpsiloFsa, Epsilo Faa	Camp, concisus
C. corcagiensis	CIT045^T	JFAP01000000	GenBank	21	Draft	1673184	31.86	1722	125	3	38	1	EpsiloFsa, Epsilo Faa	Camp
C. cuniculorum	DSM 23162^T	JHZL01000000	GenBank	46	Draft	1868556	31.24	1828	88	3	42	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
C. cuniculorum	RM 8641^T	RM8641	Dr William Miller	1	Complete	1931016	31.16	1891	90	6	40	1	EpsiloFsa, Epsilo Faa	Camp
C. curvus	525.92	CP000767	GenBank	1	Complete	1971264	44.54	1945	178	9	47	1	EpsiloFsa, Epsilo Faa	Camp
C. curvus	DSM 6644^T	AQXN01000000	GenBank	8	Draft	1954904	44.27	1920	163	4	36	1	EpsiloFsa, Epsilo Faa	Camp
C. fetus	MMM01	JRKX01000000	GenBank	143	Draft	1740393	33.12	2226	95	3	35	1	EpsiloFsa, Epsilo Faa	Camp
C. fetus subsp. *fetus*	04/554	CP008808-9	GenBank	2	Complete	1826626	33.15	1853	114	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. fetus subsp. *fetus*	82-40	CP000487	GenBank	1	Complete	1773615	33.31	1739	112	9	44	1	EpsiloFsa, Epsilo Faa	Camp

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	conten	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
C. fetus subsp. *testudinum*	03-427^{T}	CP006833	GenBank	1	Complete	1775480	33.12	1727	109	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. fetus subsp. *venerealis*	99541	ASTK01000000	GenBank	218	Draft	1871719	33.06	1889	110	3	31	1	EpsiloFsa, Epsilo Faa	Camp
C. fetus subsp. *venerealis*	642-21	AJSG01000000	GenBank	127	Draft	1943513	33.1	1984	118	7	40	1	EpsiloFsa, Epsilo Faa	Camp
C. fetus subsp. *venerealis*	84-112	HG004426	GenBank	1	Complete	1926886	33.34	1955	118	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. fetus subsp. *venerealis*	97/608	CP008810	GenBank	1	Complete	1935028	33.31	1963	118	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. fetus subsp. *venerealis*	Azul-94	ACLG01000000	GenBank	1187	Draft	2054917	36.53	2950	68	4	42	1	EpsiloFsa, Epsilo Faa	Camp
C. fetus subsp. *venerealis*	B6	AJMC01000000	GenBank	81	Draft	1945095	33.2	1965	119	7	37	1	EpsiloFsa, Epsilo Faa	Camp
C. fetus subsp. *venerealis*	cfvi03/293	CP006999-7002	GenBank	4	Complete	1996728	33.08	2028	127	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. fetus subsp. *venerealis*	NCTC 10354^{T}	AFGH01000000	GenBank	45	Draft	1858789	33.23	2059	111	3	37	1	EpsiloFsa, Epsilo Faa	Camp
C. gracilis	RM 3268^{T}	RM3268	Dr William Miller	1	Complete	2281652	46.56	2135	216	9	43	1	EpsiloFsa, Epsilo Faa	Camp
C. gracilis	RM3268^{T}	ACYG01000000	GenBank	33	Draft	2255573	46.62	2131	210	3	39	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
C. helveticus	ACP123b	ACP123b	Dr Krunolav Bojanic	170	Draft	1902264	34.23	1998	127	4	36	1	KrunoFsa	Camp
C. helveticus	ACP141a	ACP141a	Dr Krunolav Bojanic	147	Draft	1825876	34.41	1898	113	4	40	1	KrunoFsa	Camp
C. helveticus	ACP175a	ACP175a	Dr Krunolav Bojanic	134	Draft	1832474	34.4	1922	114	4	39	1	KrunoFsa	Camp
C. helveticus	RM 3228^{T}	RM3228	Dr William Miller	1	Complete	1759741	34.62	1894	107	9	42	1	EpsiloFsa, Epsilo Faa	Camp
C. hominis	ATCC BAA-381^{T}	CP000775-6	GenBank	2	Complete	1714951	31.74	1630	94	9	45	1	EpsiloFsa, Epsilo Faa	Camp
C. hyointestinalis subsp. *hyointestinalis*	DSM 19053^{T}	JHQP01000000	GenBank	26	Draft	1733268	33.93	1845	122	3	41	1	EpsiloFsa, Epsilo Faa	Camp
C. hyointestinalis subsp. *hyointestinalis*	RM 4092	RM4092	Dr William Miller	1	Complete	1753383	33.96	1791	122	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. hyointestinalis subsp. *lawsonii*	RM 4096	RM4096	Dr William Miller	1	Complete	1753279	33.55	1800	105	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. iguaniorum	1485E^{T}	CP009043-4	GenBank	2	Complete	1754638	35.74	1793	125	9	44	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
C. insulaenigrae	NCTC 12927^{T}	CP007770	GenBank	1	Complete	1465081	28.19	1480	77	9	42	1	EpsiloFsa, Epsilo Faa	Camp
C. insulaenigrae	RM 5435^{T}	RM5435	Dr William Miller	1	Complete	1465081	28.19	1480	77	9	42	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
C. jejuni	4031	HG428754	GenBank	1	Complete	1669329	30.47	1680	118	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	04199	CAFV01000000	GenBank	71	Draft	1732222	30.4	1819	121	3	38	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	10186	AUUG01000000	GenBank	24	Draft	1658236	30.33	1735	111	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	10227	AUUI01000000	GenBank	25	Draft	1717475	30.24	1756	116	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	20176	AUUQ01000000	GenBank	30	Draft	1725883	30.24	1800	113	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	30286	AUUH01000000	GenBank	18	Draft	1645076	30.44	1697	114	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	30318	AUUJ01000000	GenBank	28	Draft	1695786	30.29	1785	111	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	32488	CP006006	GenBank	1	Complete	1702398	30.49	1743	107	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	1_12S	CCCY01000000	GenBank	48	Draft	1870123	30.59	1904	131	12	66	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0032	BIGS0032	Dryad	92	Draft	1650105	30.42	1671	111	3	38	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0034	BIGS0034	Dryad	91	Draft	1643032	30.42	1690	113	3	34	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0036	BIGS0036	Dryad	17	Draft	782109	30.14	801	49	0	7	0	EpsiloFsa, Epsilo Faa	Camp

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	conten	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
C. jejuni	BIGS0037	BIGS0037	Dryad	113	Draft	1667560	30.43	1700	115	3	42	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0039	BIGS0039	Dryad	84	Draft	1616482	30.42	1623	112	3	40	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0040	BIGS0040	Dryad	195	Draft	1726327	30.42	1771	113	3	34	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0042	BIGS0042	Dryad	41	Draft	1672737	30.47	1723	104	3	38	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0043	BIGS0043	Dryad	3226	Draft	2069686	30.38	2530	69	3	34	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0045	BIGS0045	Dryad	72	Draft	1609163	30.43	1606	113	3	37	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0048	BIGS0048	Dryad	25	Draft	901732	30.49	943	56	1	24	0	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0049	BIGS0049	Dryad	181	Draft	1663148	30.4	1693	112	3	35	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0052	BIGS0052	Dryad	102	Draft	1616567	30.58	1605	116	5	37	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0054	BIGS0054	Dryad	177	Draft	1281440	30.52	1250	91	3	21	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0055	BIGS0055	Dryad	271	Draft	1625711	30.54	1619	109	3	34	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0056	BIGS0056	Dryad	62	Draft	1607778	30.53	1605	114	6	38	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0057	BIGS0057	Dryad	158	Draft	1620325	30.48	1615	110	2	35	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0059	BIGS0059	Dryad	60	Draft	1705325	30.38	1759	118	3	37	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0060	BIGS0060	Dryad	58	Draft	1658292	30.4	1715	114	4	40	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0062	BIGS0062	Dryad	100	Draft	1658436	30.37	1677	114	4	35	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0063	BIGS0063	Dryad	139	Draft	1679349	30.43	1705	116	4	38	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0064	BIGS0064	Dryad	113	Draft	1698035	30.29	1753	113	4	37	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0065	BIGS0065	Dryad	112	Draft	1761202	30.37	1823	113	3	35	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0066	BIGS0066	Dryad	129	Draft	1776210	30.27	1808	115	3	35	0	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0067	BIGS0067	Dryad	143	Draft	1688706	30.26	1750	110	3	34	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0068	BIGS0068	Dryad	251	Draft	1838022	30.06	1909	118	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0069	BIGS0069	Dryad	144	Draft	1821236	30.06	1900	120	3	37	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0070	BIGS0070	Dryad	343	Draft	1595762	30.54	1565	108	3	34	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0071	BIGS0071	Dryad	84	Draft	1689713	30.42	1745	113	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0072	BIGS0072	Dryad	63	Draft	1692341	30.39	1751	115	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0073	BIGS0073	Dryad	102	Draft	1651079	30.46	1700	113	3	34	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0074	BIGS0074	Dryad	65	Draft	1606379	30.49	1650	112	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0075	BIGS0075	Dryad	76	Draft	1693845	30.38	1747	115	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0077	BIGS0077	Dryad	65	Draft	1692435	30.39	1744	115	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0078	BIGS0078	Dryad	67	Draft	1693941	30.39	1755	115	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0079	BIGS0079	Dryad	120	Draft	1645238	30.44	1669	110	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0080	BIGS0080	Dryad	73	Draft	1734017	30.19	1810	119	3	35	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0081	BIGS0081	Dryad	106	Draft	1617200	30.46	1645	110	3	36	1	EpsiloFsa, Epsilo Faa	Camp

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	conten	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
C. jejuni	BIGS0082	BIGS0082	Dryad	163	Draft	1631119	30.63	1603	115	3	37	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0083	BIGS0083	Dryad	617	Draft	1667030	30.55	1657	101	3	34	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0084	BIGS0084	Dryad	58	Draft	1605483	30.47	1625	115	3	37	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0085	BIGS0085	Dryad	93	Draft	1654563	30.51	1691	111	3	34	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0086	BIGS0086	Dryad	197	Draft	1659558	30.4	1674	112	3	35	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0087	BIGS0087	Dryad	62	Draft	1715399	30.32	1775	110	3	38	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0088	BIGS0088	Dryad	72	Draft	1716038	30.2	1765	110	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0089	BIGS0089	Dryad	130	Draft	1659711	30.42	1700	113	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0090	BIGS0090	Dryad	64	Draft	1603131	30.48	1612	114	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0091	BIGS0091	Dryad	47	Draft	1602224	30.47	1607	109	3	38	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0092	BIGS0092	Dryad	98	Draft	1607691	30.44	1607	111	3	37	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0094	BIGS0094	Dryad	82	Draft	1762939	30.38	1838	120	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0096	BIGS0096	Dryad	922	Draft	1894874	30.31	1833	105	3	37	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0097	BIGS0097	Dryad	114	Draft	1629692	30.48	1652	116	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0099	BIGS0099	Dryad	64	Draft	1720953	30.36	1777	114	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0100	BIGS0100	Dryad	192	Draft	1626334	30.83	1599	114	6	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0102	BIGS0102	Dryad	63	Draft	1607515	30.44	1614	114	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0103	BIGS0103	Dryad	70	Draft	1641790	30.45	1662	114	3	38	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0104	BIGS0104	Dryad	115	Draft	1654338	30.48	1669	118	3	35	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0105	BIGS0105	Dryad	75	Draft	1697013	30.26	1765	113	3	35	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0106	BIGS0106	Dryad	58	Draft	1714044	30.31	1787	115	3	37	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0107	BIGS0107	Dryad	273	Draft	1838061	30.04	1918	114	3	37	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0108	BIGS0108	Dryad	80	Draft	1743461	30.33	1824	116	3	38	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0109	BIGS0109	Dryad	152	Draft	1759789	30.2	1830	115	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0110	BIGS0110	Dryad	72	Draft	1656261	30.36	1682	115	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0111	BIGS0111	Dryad	80	Draft	1649834	30.45	1674	117	3	37	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0112	BIGS0112	Dryad	55	Draft	1618948	30.45	1607	117	3	37	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0113	BIGS0113	Dryad	72	Draft	1665144	30.42	1711	112	3	38	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0114	BIGS0114	Dryad	79	Draft	1633592	30.5	1652	115	3	35	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0116	BIGS0116	Dryad	65	Draft	1656471	30.36	1679	115	3	38	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0117	BIGS0117	Dryad	113	Draft	1697696	30.41	1749	111	3	38	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0119	BIGS0119	Dryad	66	Draft	1621668	30.44	1609	117	3	38	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0122	BIGS0122	Dryad	52	Draft	1582720	30.42	1581	98	3	38	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0124	BIGS0124	Dryad	112	Draft	1603669	30.47	1586	109	3	34	1	EpsiloFsa, Epsilo Faa	Camp

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	conten	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
C. jejuni	BIGS0125	BIGS0125	Dryad	61	Draft	1578916	30.43	1572	106	3	39	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0126	BIGS0126	Dryad	149	Draft	1663834	30.28	1693	115	5	34	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0127	BIGS0127	Dryad	2598	Draft	1538093	31.32	2254	38	3	26	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0128	BIGS0128	Dryad	53	Draft	1600143	30.47	1607	116	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0129	BIGS0129	Dryad	143	Draft	1561449	30.42	1577	98	3	37	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0130	BIGS0130	Dryad	108	Draft	1669953	30.12	1723	114	3	37	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BIGS0131	BIGS0131	Dryad	60	Draft	1616162	30.49	1623	117	3	37	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	BJ-CJD101	ARWV01000000	GenBank	22	Draft	1610177	30.48	1629	108	3	37	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	cj1	AUUL01000000	GenBank	52	Draft	1752906	30.18	1780	115	3	38	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	cj2	AUUM01000000	GenBank	48	Draft	1677685	30.39	1706	112	2	39	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	cj3	AUUN01000000	GenBank	44	Draft	1732132	30.34	1797	113	3	38	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	cj5	AUUK01000000	GenBank	29	Draft	1631773	30.34	1689	108	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	CJJ5070	CCXG01000000	GenBank	15	Draft	1651321	30.33	1735	107	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	G1	JRLT01000000	GenBank	14	Draft	1721530	30.5	1838	115	8	39	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	g113	AQPK01000000	GenBank	17	Draft	1598550	30.54	1643	114	3	38	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
C. jejuni	HB93-13	AANQ01000000	GenBank	35	Draft	1694788	30.59	1686	115	12	47	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	K1	AUUO01000000	GenBank	18	Draft	1619749	30.5	1672	114	3	38	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	K5	AUUP01000000	GenBank	466	Draft	3303828	30.71	3357	215	3	55	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	Le_204R	CCDB01000000	GenBank	48	Draft	1680872	30.71	1671	110	18	50	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	Le_755	CCDC01000000	GenBank	60	Draft	1684485	30.7	1669	110	18	50	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	Ma_1	CCCZ01000000	GenBank	54	Draft	1663694	30.58	1662	110	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	Ma_B	CCDD01000000	GenBank	57	Draft	1668359	30.62	1667	110	12	46	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	NCCP No. 15742	APJU01000000	GenBank	5	Draft	1592587	30.56	1666	108	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	Po_1	CCDA01000000	GenBank	73	Draft	1672030	30.71	1671	110	15	48	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	Po_2	CCDE01000000	GenBank	50	Draft	1674783	30.66	1666	110	15	48	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	RM1221	CP000025	GenBank	1	Complete	1777831	30.31	1871	112	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	X	AVFM02000000	GenBank	85	Draft	1731223	30.49	1911	110	9	40	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni	255	ARWS01000000	GenBank	25	Draft	1634595	30.28	1677	114	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *doylei*	269.97	CP000768	GenBank	1	Complete	1845106	30.57	1979	98	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *doylei*	ATCC 49349	JNJR01000000	GenBank	46	Draft	1732618	30.49	1854	94	6	40	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	260.94	AANK01000000	GenBank	10	Draft	1657846	30.54	1681	114	7	45	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	305	ADHL01000000	GenBank	333	Draft	1808274	30.4	1962	96	8	39	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	327	ADHM01000000	GenBank	48	Draft	1618613	30.54	1737	111	5	40	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	414	ADGM01000000	GenBank	35	Draft	1686453	29.8	1800	105	3	34	1	EpsiloFsa, Epsilo Faa	Camp

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	conten	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
C. jejuni subsp. *jejuni*	1213	AIPG01000000	GenBank	125	Draft	1648061	30.53	1655	106	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	1336	ADGL01000000	GenBank	35	Draft	1674540	30.2	1749	106	3	40	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	1577	AIPH01000000	GenBank	170	Draft	1703661	30.32	1756	108	3	31	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	1798	AIPI01000000	GenBank	106	Draft	1602805	30.4	1637	102	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	1854	AIPJ01000000	GenBank	110	Draft	1622514	30.4	1634	106	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	1893	AIPK01000000	GenBank	101	Draft	1626431	30.42	1675	107	3	34	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	1928	AIPL01000000	GenBank	127	Draft	1637403	30.53	1670	103	3	32	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	04197	CAFU01000000	GenBank	83	Draft	1701336	30.44	1809	113	3	39	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	6399	CAFT01000000	GenBank	43	Draft	1627034	30.45	1654	110	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	51037	AIPB01000000	GenBank	207	Draft	1747358	30.21	1774	97	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	51494	AINZ01000000	GenBank	215	Draft	1803640	30.17	1867	104	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	53161	AION01000000	GenBank	123	Draft	1562744	30.5	1601	95	3	34	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	55037	AIOH01000000	GenBank	104	Draft	1589716	30.5	1587	106	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	60004	AIOE01000000	GenBank	143	Draft	1675414	30.31	1715	107	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	81116	CP000814	GenBank	1	Complete	1628115	30.54	1617	115	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	86605	AIOJ01000000	GenBank	120	Draft	1638403	30.42	1661	106	2	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	87330	AIPD01000000	GenBank	107	Draft	1610401	30.45	1627	108	3	34	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	87459	AIPE01000000	GenBank	201	Draft	1727997	30.39	1779	97	3	32	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	00-2425	CP006729	GenBank	1	Complete	1718982	30.51	1762	115	9	40	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	00-2426	CP006708	GenBank	1	Complete	1680813	30.52	1708	115	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	00-2538	CP006707	GenBank	1	Complete	1719369	30.51	1762	115	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	00-2544	CP006709-10	GenBank	2	Complete	1766434	30.48	1816	119	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	110-21	AIPC01000000	GenBank	113	Draft	1619234	30.51	1638	109	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	129-258	AINY01000000	GenBank	102	Draft	1608162	30.56	1623	98	3	32	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	140-16	AIPF01000000	GenBank	123	Draft	1678201	30.31	1711	105	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	1997-1	AIOT01000000	GenBank	112	Draft	1604303	30.4	1627	105	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	1997-10	AIOY01000000	GenBank	193	Draft	1780746	30.21	1821	107	3	31	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	1997-11	AIOZ01000000	GenBank	115	Draft	1601760	30.49	1597	101	3	34	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	1997-14	AIPA01000000	GenBank	196	Draft	1766943	30.29	1816	99	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	1997-4	AIOW01000000	GenBank	129	Draft	1670223	30.38	1691	96	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	1997-7	AIOX01000000	GenBank	76	Draft	1589764	30.39	1643	102	3	32	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	2008-1025	AIOP01000000	GenBank	139	Draft	1661042	30.42	1686	109	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	2008-831	AIOV01000000	GenBank	120	Draft	1609412	30.46	1623	108	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	2008-872	AIOR01000000	GenBank	114	Draft	1602313	30.4	1628	101	3	33	1	EpsiloFsa, Epsilo Faa	Camp

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	conten	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
C. jejuni subsp. *jejuni*	2008-894	AIOQ01000000	GenBank	116	Draft	1627929	30.36	1629	96	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	2008-979	AIOU01000000	GenBank	217	Draft	1798433	30.23	1883	100	2	32	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	2008-988	AIOS01000000	GenBank	206	Draft	1820783	30.25	1883	115	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	81-176	AASL01000000	GenBank	1	Draft	1616175	30.62	1727	103	9	40	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
C. jejuni subsp. *jejuni*	81-176	CP000538, 49-50	GenBank	3	Complete	1699052	30.48	1723	116	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	81-176-DRH212	AZNT01000000	GenBank	24	Draft	1412401	30.33	1689	96	1	34	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
C. jejuni subsp. *jejuni*	81-176-UNCW7	AZNS01000000	GenBank	34	Draft	1322824	30.48	1356	90	5	38	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
C. jejuni subsp. *jejuni*	81-176-UNCW9	AZQV01000000	GenBank	38	Draft	1662527	30.52	1759	110	9	45	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
C. jejuni subsp. *jejuni*	84-25	AANT02000000	GenBank	5	Draft	1671624	30.44	1724	114	5	40	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	ATCC 33560^{T}	AIJN01000000	GenBank	43	Draft	1732191	30.18	1897	98	3	40	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	ATCC 33560^{T}	AIOL01000000	GenBank	220	Draft	1712461	30.19	1748	99	3	32	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	BH-01-0142	ABKD01000000	GenBank	53	Draft	118623	27.68	149	3	3	11	0	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	BIGS0004	ANGO01000000	GenBank	108	Draft	1596969	30.66	1617	106	2	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	BIGS0013	ANGX01000000	GenBank	86	Draft	1551351	30.65	1599	108	3	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	BIGS0014	ANGY01000000	GenBank	324	Draft	1543877	30.86	1487	93	3	29	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	BIGS0022	ANHG01000000	GenBank	169	Draft	1573736	30.85	1615	111	3	32	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	BIGS0030	ANHK01000000	GenBank	422	Draft	1607358	30.36	1587	90	1	30	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	CF93-6	AANJ01000000	GenBank	14	Draft	1676304	30.47	1727	113	5	40	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	CG8421	ABGQ01000000	GenBank	20	Draft	1608937	30.37	1734	102	0	32	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
C. jejuni subsp. *jejuni*	CG8421	CP005388	GenBank	1	Complete	1636029	30.51	1691	107	8	44	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	CG8486	AASY01000000	GenBank	19	Draft	1597692	30.44	1814	103	0	37	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	D2600	AGTF01000000	GenBank	56	Draft	1622329	30.46	1676	110	4	38	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	DFVF1099	ADHK01000000	GenBank	71	Draft	1733857	30.4	1887	111	5	43	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	H22082	AEIP01000000	GenBank	28	Draft	1659123	30.58	1678	111	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	HN-CJD07035	ARYE01000000	GenBank	13	Draft	1652971	30.47	1667	117	2	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	IA3902	CP001876-7	GenBank	2	Complete	1672219	30.46	1700	118	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	ICDCCJ07001	CP002029-30	GenBank	2	Complete	1708924	30.55	1848	115	9	45	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	ICDCCJ07002	APNP01000000	GenBank	14	Draft	1694682	30.45	1726	117	2	38	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	ICDCCJ07004	APNQ01000000	GenBank	73	Draft	1695092	30.5	1722	114	3	38	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	JCM 2013	BALI01000000	GenBank	48	Draft	1662874	30.39	1769	121	3	40	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	LMG 23210	AIPN01000000	GenBank	164	Draft	1764103	30.2	1819	114	3	32	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	LMG 23211	AIPO01000000	GenBank	112	Draft	1672359	30.42	1702	102	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	LMG 23216	AIOA01000000	GenBank	121	Draft	1474288	30.34	1483	85	3	32	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	LMG 23218	AIOB01000000	GenBank	106	Draft	1678928	30.45	1697	105	3	33	1	EpsiloFsa, Epsilo Faa	Camp

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	conten	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
C. jejuni subsp. *jejuni*	LMG 23223	AIOC01000000	GenBank	105	Draft	1632500	30.43	1703	103	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	LMG 23263	AIOD01000000	GenBank	172	Draft	1744709	30.27	1799	106	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	LMG 23264	AIOF01000000	GenBank	147	Draft	1720846	30.35	1783	105	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	LMG 23269	AIOG01000000	GenBank	124	Draft	1628395	30.47	1680	98	3	32	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	LMG 23357	AIOK01000000	GenBank	126	Draft	1672858	30.26	1695	105	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	LMG 9081	AIOM01000000	GenBank	108	Draft	1593841	30.48	1616	104	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	LMG 9217	AIOO01000000	GenBank	118	Draft	1653445	30.37	1708	107	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	LMG 9872	AIPM01000000	GenBank	91	Draft	1620398	30.41	1674	101	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	LMG 9879	AIOI01000000	GenBank	129	Draft	1650812	30.47	1695	110	3	33	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	M1	CP001900	GenBank	1	Complete	1616648	30.6	1635	119	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	MTVDSCj20	CP008787	GenBank	1	Complete	1651739	30.52	1665	122	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	NCTC 11168	AL111168	GenBank	1	Complete	1641481	30.55	1658	114	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	NCTC 11168-BN148	HE978252	GenBank	1	Complete	1641481	30.55	1658	114	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	NCTC 11168-GSv	CP006689	GenBank	1	Complete	1641482	30.55	1660	114	9	44	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
C. jejuni subsp. *jejuni*	NCTC 11168-K12E5	CP006685	GenBank	1	Complete	1641481	30.55	1656	114	9	44	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
C. jejuni subsp. *jejuni*	NCTC 11168-Kf1	CP006686	GenBank	1	Complete	1641480	30.55	1657	114	9	44	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
C. jejuni subsp. *jejuni*	NCTC 11168-mcK12E5	CP006687	GenBank	1	Complete	1641471	30.55	1658	114	9	44	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
C. jejuni subsp. *jejuni*	NCTC 11168-mfK12E5	CP006688	GenBank	1	Complete	1641469	30.55	1656	114	9	44	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
C. jejuni subsp. *jejuni*	NW	AGTE01000000	GenBank	60	Draft	1652691	30.31	1712	110	3	40	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	P110B	AEIO01000000	GenBank	29	Draft	1656341	30.57	1670	110	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	P854	AKFN01000000	GenBank	185	Draft	1744857	30.09	1820	113	3	34	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	PT14	CP003871	GenBank	1	Complete	1635252	30.54	1644	113	9	41	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	R14	CP005081	GenBank	1	Complete	1795858	30.41	1954	108	9	44	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
C. jejuni subsp. *jejuni*	RB922	CAFS01000000	GenBank	156	Draft	1715324	30.45	1764	111	4	35	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	S3	CP001960-1	GenBank	2	Complete	1724586	30.45	1809	112	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	xy259	CAFR01000000	GenBank	68	Draft	1709783	30.34	1770	123	4	36	1	EpsiloFsa, Epsilo Faa	Camp
C. jejuni subsp. *jejuni*	YH001	CP010058	GenBank	1	Complete	1712361	30.54	1821	109	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. lanienae	RM 3663^{T}	RM3663	Dr William Miller	1	Complete	1594551	34.6	1596	76	9	40	1	EpsiloFsa, Epsilo Faa	Camp
C. lari	RM 16704	RM16704	Dr William Miller	1	Complete	1557535	28.47	1548	103	9	46	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
C. lari	RM2100	CP000932-3	GenBank	2	Complete	1571661	29.62	1576	107	9	46	1	EpsiloFsa, Epsilo Faa	Camp, lari
C. lari subsp. *concheus*	LMG 11760	CP007771	GenBank	1	Complete	1502102	29.73	1494	99	9	43	1	EpsiloFsa, Epsilo Faa	Camp, lari
C. lari subsp. *concheus*	RM 2825	RM2825	Dr William Miller	1	Complete	1502102	29.73	1494	99	9	43	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
C. mucosalis	DSM 21682^{T}	JHQQ01000000	GenBank	42	Draft	1751734	36.64	1914	102	4	41	1	EpsiloFsa, Epsilo Faa	Camp
C. mucosalis	RM 4113	RM4113	Dr William Miller	1	Complete	1860004	36.54	2002	112	6	43	1	EpsiloFsa, Epsilo Faa	Camp

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	content	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
C. peloridis	LMG 23910^T	CP007766-8	GenBank	2	Complete	1762793	28.46	1683	111	9	46	1	EpsiloFsa, Epsilo Faa	Camp
C. peloridis	RM 14092^T	RM14092	Dr William Miller	1	Complete	1695385	28.5	1618	107	9	46	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
C. rectus	ATCC 33238^T	RM6916	Dr William Miller	1	Complete	2571691	44.72	2329	201	9	49	1	EpsiloFsa, Epsilo Faa	Camp
C. rectus	RM3267	ACFU01000000	GenBank	89	Draft	2513107	44.85	2327	198	3	41	1	EpsiloFsa, Epsilo Faa	Camp
C. showae	CC57C	AOTD01000000	GenBank	274	Draft	2192767	45.36	2202	198	3	30	1	EpsiloFsa, Epsilo Faa	Camp
C. showae	CSUNSWCD	AMZQ01000000	GenBank	23	Draft	2125173	45.13	2137	197	3	39	1	EpsiloFsa, Epsilo Faa	Camp
C. showae	RM 3277^T	RM3277	Dr William Miller	1	Complete	2097888	45.66	2005	207	9	48	1	EpsiloFsa, Epsilo Faa	Camp
C. showae	RM3277	ACVQ01000000	GenBank	33	Draft	2072007	45.69	1990	203	3	40	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
C. sputorum	INTA08/209	JMTI01000000	GenBank	51	Draft	1780580	29.14	1842	81	2	43	1	EpsiloFsa, Epsilo Faa	Camp
C. sputorum biovar *fecalis*	RM 4121	RM4121	Dr William Miller	1	Complete	1757291	29.72	1768	97	9	47	1	EpsiloFsa, Epsilo Faa	Camp
C. sputorum biovar *paraureolyticus*	RM 4120	RM4120	Dr William Miller	1	Complete	1725014	29.63	1731	97	9	47	1	EpsiloFsa, Epsilo Faa	Camp
C. sputorum biovar *sputorum*	RM 3237	RM3237	Dr William Miller	1	Complete	1752258	29.71	1758	100	9	47	1	EpsiloFsa, Epsilo Faa	Camp
C. subantarcticus	LMG 24374	CP007772	GenBank	1	Complete	1782536	29.94	1754	122	9	46	1	EpsiloFsa, Epsilo Faa	Camp
C. subantarcticus	LMG 24377^T	CP007773	GenBank	1	Complete	1852995	29.75	1864	127	9	46	1	EpsiloFsa, Epsilo Faa	Camp
C. subantarcticus	RM 8521	RM8521	Dr William Miller	1	Complete	1782536	29.94	1754	122	9	46	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
C. subantarcticus	RM 8523^T	RM8523	Dr William Miller	1	Complete	1852995	29.75	1864	127	9	46	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
C. upsaliensis	ACP170b	ACP170b	Dr Krunolav Bojanic	125	Draft	1778375	34.29	1834	117	6	40	1	KrunoFsa	Camp
C. upsaliensis	ACP5b	ACP5b	Dr Krunolav Bojanic	106	Draft	1722053	34.7	1756	105	3	38	1	KrunoFsa	Camp
C. upsaliensis	BD16E4a	BD16E4a	Dr Krunolav Bojanic	110	Draft	1725260	34.59	1752	100	6	39	1	KrunoFsa	Camp
C. upsaliensis	DSM 5365^T	JHZN01000000	GenBank	44	Draft	1619002	34.96	1629	81	3	42	1	EpsiloFsa, Epsilo Faa	Camp
C. upsaliensis	JV21	AEPU01000000	GenBank	49	Draft	1621689	34.8	1663	95	3	40	1	EpsiloFsa, Epsilo Faa	Camp
C. upsaliensis	RM3195	AAFJ01000000	GenBank	20	Draft	1773834	34.27	1847	102	5	44	1	EpsiloFsa, Epsilo Faa	Camp
C. ureolyticus	ACS-301-V-Sch3b	AGYD01000000	GenBank	19	Draft	1691524	29.33	1723	114	6	42	1	EpsiloFsa, Epsilo Faa	Camp
C. ureolyticus	CIT007	JFJK01000000	GenBank	25	Draft	1665702	29.02	1680	116	4	41	1	EpsiloFsa, Epsilo Faa	Camp
C. ureolyticus	DSM 20703^T	ARGD01000000	GenBank	37	Draft	1740135	28.92	1770	106	5	41	1	EpsiloFsa, Epsilo Faa	Camp
C. ureolyticus	RM 4126	RM4126	Dr William Miller	1	Complete	1641887	29.23	1654	103	9	44	1	EpsiloFsa, Epsilo Faa	Camp
C. volucris	LMG 24379	CP007774	GenBank	1	Complete	1517949	28.57	1512	96	9	43	1	EpsiloFsa, Epsilo Faa	Camp
C. volucris	RM 9725	RM9725	Dr William Miller	1	Complete	1517949	28.57	1512	96	9	43	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
"Urease positive thermophilic *Campylobacter* (U	CCUG 22395	CP007776	GenBank	1	Complete	1523028	29.86	1512	109	9	46	1	EpsiloFsa, Epsilo Faa	Camp, lari
"Urease positive thermophilic *Campylobacter* (U	NCTC 11845	CP007775	GenBank	1	Complete	1791508	29.36	1753	121	9	46	1	EpsiloFsa, Epsilo Faa	Camp, lari
"Urease positive thermophilic *Campylobacter* (U	RM 16701	CP007777	GenBank	1	Complete	1516455	29.9	1494	109	9	43	1	EpsiloFsa, Epsilo Faa	Camp, lari
"Urease positive thermophilic *Campylobacter* (U	RM 16701	RM16701	Dr William Miller	1	Complete	1516455	29.9	1494	109	9	43	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
"Urease positive thermophilic *Campylobacter* (U	RM 16712	CP007778	GenBank	1	Complete	1565009	29.74	1544	104	9	43	1	EpsiloFsa, Epsilo Faa	Camp, lari
"Urease positive thermophilic *Campylobacter* (U	RM 16712	RM16712	Dr William Miller	1	Complete	1565009	29.74	1544	104	9	43	1	EpsiloFsa, Epsilo Faa	Camp (ignored)

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	conten	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
"Urease positive thermophilic *Campylobacter* (U	RM 3659	RM3659	Dr William Miller	1	Complete	1791508	29.36	1753	121	9	46	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
"Urease positive thermophilic *Campylobacter* (U	RM 4110	RM4110	Dr William Miller	1	Complete	1523028	29.86	1512	109	9	46	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
Campylobacter sp.	10_1_50	ACWJ01000000	GenBank	32	Draft	1890008	39.55	1905	155	3	43	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
Campylobacter sp.	B1491	B1491	Dr Patrick Biggs	25	Draft	1647693	27.44	1659	65	3	35	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
Campylobacter sp.	B1821b	B1821b	Dr Patrick Biggs	179	Draft	1979432	29.97	2022	115	3	35	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
Campylobacter sp.	B1932a	B1932a	Dr Patrick Biggs	270	Draft	2307250	29.83	2387	152	3	36	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
Campylobacter sp.	B2020a	B2020a	Dr Patrick Biggs	262	Draft	1835080	31.61	1893	111	3	34	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
Campylobacter sp.	B2092c	B2092c	Dr Patrick Biggs	44	Draft	1870806	31.82	1790	124	0	32	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
Campylobacter sp.	B2098b	B2098b	Dr Patrick Biggs	28	Draft	1946889	31.78	1938	108	6	41	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
Campylobacter sp.	FOBRC14	ALJJ01000000	GenBank	19	Draft	2142354	44.09	2205	186	4	39	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
Campylobacter sp.	J148a	J148a	Dr Patrick Biggs	236	Draft	2230656	29.93	2251	147	3	33	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
Campylobacter sp.	MIT 97-5078	JRMR01000000	GenBank	184	Draft	2454906	33.96	2445	168	8	40	2	EpsiloFsa, Epsilo Faa	Camp (ignored)
Campylobacter sp.	RM 16704	CP007769	GenBank	1	Complete	1557536	28.47	1543	103	9	46	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
Campylobacter sp.	W441b	W441b	Dr Patrick Biggs	1536	Draft	1407036	27.88	1524	62	0	0	0	EpsiloFsa, Epsilo Faa	Camp (ignored)
Campylobacter sp.	W677a	W677a	Dr Patrick Biggs	37	Draft	1639146	27.48	1645	70	3	51	1	EpsiloFsa, Epsilo Faa	Camp (ignored)
Caminibacter mediatlanticus	TB-2^T	ABCJ01000000	GenBank	35	Draft	1663618	27.13	1729	50	18	52	1	EpsiloFsa, Epsilo Faa	
H. acinonychis	Sheeba	AM260522-3	GenBank	2	Complete	1557588	38.17	1546	87	6	36	1	EpsiloFsa, Epsilo Faa	Heli
"H. apodemus"	MIT 03-7007	JRPC01000000	GenBank	628	Draft	2114943	32.99	2086	77	4	37	0	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. bilis	ATCC 43879	ACDN02000000	GenBank	77	Draft	2523721	34.79	2349	164	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. bilis	ATCC 49314	JRPI01000000	GenBank	97	Draft	2505915	34.81	2405	147	6	38	1	EpsiloFsa, Epsilo Faa	Heli
H. bilis	ATCC 49320	JRPJ01000000	GenBank	82	Draft	2549041	34.81	2425	143	6	38	1	EpsiloFsa, Epsilo Faa	Heli
H. bilis	ATCC 51630^T	JMKW01000000	GenBank	131	Draft	2474805	34.93	2290	160	9	40	1	EpsiloFsa, Epsilo Faa	Heli
H. bilis	ATCC 51630^T	JRPG01000000	GenBank	542	Draft	2485594	34.85	2252	149	3	36	0	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. bilis	Missouri	JRPH01000000	GenBank	467	Draft	2403907	34.91	2247	146	3	35	1	EpsiloFsa, Epsilo Faa	Heli
H. bilis	WiWa	AQFW01000000	GenBank	33	Draft	2548580	34.83	2365	167	9	40	1	EpsiloFsa, Epsilo Faa	Heli
H. bizzozeronii	CCUG 35545^T	CAGP01000000	GenBank	147	Draft	1775627	45.78	2373	79	4	37	1	EpsiloFsa, Epsilo Faa	Heli
H. bizzozeronii	CIII-1	FR871757-8	GenBank	2	Complete	1798412	45.9	1870	86	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. canadensis	MIT 98-5491^T	ABQS01000000	GenBank	126	Draft	1605969	33.66	1550	61	3	37	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. canadensis	MIT 98-5491^T	ACSF01000000	GenBank	4	Draft	1622996	33.7	1575	70	9	41	1	EpsiloFsa, Epsilo Faa	Heli
H. canis	NCTC 12740	AZJJ01000000	GenBank	7	Draft	1924983	45	1791	119	6	42	1	EpsiloFsa, Epsilo Faa	Heli
H. cetorum	MIT 00-7128	CP003479	GenBank	1	Complete	1947646	34.53	1766	141	6	39	1	EpsiloFsa, Epsilo Faa	Heli
H. cetorum	MIT 99-5656^T	CP003481	GenBank	1	Complete	1833666	35.56	1710	90	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. cinaedi	ATCC BAA-847^T	AP012492	GenBank	1	Complete	2240130	38.34	2325	123	6	40	1	EpsiloFsa, Epsilo Faa	Heli
H. cinaedi	CCUG 18818^T	ABQT01000000	GenBank	96	Draft	2182709	38.46	2251	116	3	38	1	EpsiloFsa, Epsilo Faa	Heli

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	conten	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
H. cinaedi	PAGU611	AP012344-5	GenBank	2	Complete	2101402	38.55	2140	108	6	40	1	EpsiloFsa, Epsilo Faa	Heli
H. felis	ATCC 49179^T	NC_014810.1	GenBank	1	Complete	1672681	44.51	1668	78	5	37	1	EpsiloFsa, Epsilo Faa	Heli
H. fennelliae	MRY12-0050	BASD01000000	GenBank	49	Draft	2155647	37.9	2121	127	3	38	1	EpsiloFsa, Epsilo Faa	Heli
H. heilmannii	ASB1^T	HE984298-9	GenBank	2	Complete	1806315	47.71	1826	86	7	42	1	EpsiloFsa, Epsilo Faa	Heli
H. hepaticus	ATCC 51449	AE017125	GenBank	1	Complete	1799146	35.93	1798	104	3	38	1	EpsiloFsa, Epsilo Faa	Heli
"H. jaachi"	MIT 09-6949^T	JRPR01000000	GenBank	140	Draft	1905102	41.04	1853	127	4	39	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. macacae	MIT 99-5501^T	AZJI01000000	GenBank	12	Draft	2358757	40.6	1941	134	6	39	1	EpsiloFsa, Epsilo Faa	Heli
"H. magdeburgensis"	MIT 96-1001	JRPE01000000	GenBank	360	Draft	2084456	38.75	2067	125	3	38	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. muridarum	ST1^T	JRPD01000000	GenBank	92	Draft	2352062	32.69	2116	66	3	46	1	EpsiloFsa, Epsilo Faa	Heli
H. mustelae	12198^T	FN555004	GenBank	1	Complete	1578097	42.47	1422	60	6	39	1	EpsiloFsa, Epsilo Faa	Heli
H. pametensis	ATCC 51478^T	JADE01000000	GenBank	13	Draft	1433836	40.11	1367	61	8	39	1	EpsiloFsa, Epsilo Faa	Heli
H. pullorum	MIT 98-5489	ABQU01000000	GenBank	131	Draft	1919070	34.15	1916	90	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	51	CP000012	GenBank	1	Complete	1589954	38.77	1512	89	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	52	CP001680	GenBank	1	Complete	1568826	38.94	1508	93	6	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	83	CP002605	GenBank	1	Complete	1617426	38.72	1558	89	6	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	908	CP002184	GenBank	1	Complete	1549666	39.3	1509	91	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	2017	CP002571	GenBank	1	Complete	1548238	39.3	1501	93	3	37	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	2018	CP002572	GenBank	1	Complete	1562832	39.29	1506	94	3	37	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	26695	AE000511	GenBank	1	Complete	1667867	38.87	1576	94	7	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	26695	CP003904	GenBank	1	Complete	1667892	38.87	1574	95	7	36	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	1089/03	JSUY01000000	GenBank	62	Draft	1587129	39.14	1506	86	3	34	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	1152/04	JSUZ01000000	GenBank	59	Draft	1588506	39.13	1504	89	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	1198/04	JSXT01000000	GenBank	57	Draft	1618315	38.95	1490	93	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	173/00	JSXX01000000	GenBank	79	Draft	1572960	39.11	1461	86	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	1786/05	JSXW01000000	GenBank	41	Draft	1619653	38.95	1525	90	4	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	1846/05	JSXV01000000	GenBank	58	Draft	1641829	38.96	1534	87	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	207/99	JSXU01000000	GenBank	47	Draft	1544379	39.2	1429	89	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	228/99	JSXY01000000	GenBank	54	Draft	1619591	39.03	1517	86	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	26695-1	AP013354	GenBank	1	Complete	1667638	38.87	1566	95	7	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	26695-1CH	AP013355	GenBank	1	Complete	1667302	38.87	1567	95	7	36	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	26695-1CL	AP013356	GenBank	1	Complete	1667239	38.87	1567	95	7	36	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	35A	CP002096	GenBank	1	Complete	1566655	38.87	1497	87	6	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	499/02	JTDG01000000	GenBank	73	Draft	1666733	38.9	1553	95	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	655/99	JSXB01000000	GenBank	45	Draft	1615595	38.95	1489	86	4	37	1	EpsiloFsa, Epsilo Faa	Heli

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	content	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
H. pylori	8A3	CADD01000000	GenBank	44	Draft	1547179	38.96	1582	79	3	37	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	98-10	ABSX01000000	GenBank	51	Draft	1571772	38.76	1537	89	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	A45	AMYU01000000	GenBank	33	Draft	1643927	38.67	1555	86	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Aklavik117	CP003483-5	GenBank	3	Complete	1636125	38.73	1530	84	6	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Aklavik86	CP003476-8	GenBank	3	Complete	1507930	39.21	1411	91	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	B128	ABSY01000000	GenBank	73	Draft	1649221	38.77	1745	85	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	B38	FM991728	GenBank	1	Complete	1576758	39.16	1499	83	7	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	B45	AFAO01000000	GenBank	63	Draft	1602587	39.04	1566	84	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	B8	FN598874	GenBank	1	Complete	1673997	38.79	1572	99	6	36	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	BCS100H1	CADC01000000	GenBank	31	Draft	1548400	38.95	1537	86	3	37	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	BM012A	CP006888	GenBank	1	Complete	1660425	38.88	1572	91	7	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	BM012B	CP007605	GenBank	1	Complete	1659060	38.88	1570	92	7	37	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	BM012S	CP006889	GenBank	1	Complete	1660469	38.88	1576	90	7	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	BM013A	CP007604	GenBank	1	Complete	1604233	38.96	1503	87	7	36	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	BM013B	CP007606	GenBank	1	Complete	1604212	38.96	1504	87	7	36	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	CCHI 33	AOTU01000000	GenBank	11	Draft	1659327	39.19	1541	104	4	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	CG-IMSS-2012	AWUL01000000	GenBank	45	Draft	1599050	39.02	1580	81	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Col2025	JOKW01000000	GenBank	54	Draft	1637839	38.98	1529	99	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	CPY1124	AKNJ01000000	GenBank	13	Draft	1563197	38.94	1536	96	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	CPY1313	AKNK01000000	GenBank	5	Draft	1581437	38.79	1550	96	5	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	CPY1662	AOTT01000000	GenBank	9	Draft	1595824	38.76	1529	85	7	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	CPY1962	AKNL01000000	GenBank	8	Draft	1561561	38.87	1516	87	4	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	CPY3281	AKNM01000000	GenBank	8	Draft	1606528	38.73	1607	89	4	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	CPY6081	AKNN01000000	GenBank	9	Draft	1599100	38.7	1560	95	5	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	CPY6261	AKNO01000000	GenBank	5	Draft	1608996	38.77	1555	90	5	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	CPY6271	AKNP01000000	GenBank	8	Draft	1602120	38.76	1531	89	5	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	CPY6311	AKNQ01000000	GenBank	9	Draft	1596646	38.75	1543	93	2	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Cuz20	CP002076	GenBank	1	Complete	1635449	38.86	1534	92	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	D33	ANIO01000000	GenBank	64	Draft	1545148	38.89	1469	90	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	E48	AYHQ01000000	GenBank	61	Draft	1661917	38.66	1839	95	2	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	ELS37	CP002953-4	GenBank	2	Complete	1669876	38.88	1560	89	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	F16	AP011940	GenBank	1	Complete	1575399	38.88	1507	87	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	F30	AP011941-2	GenBank	2	Complete	1579693	38.8	1496	89	6	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	F32	AP011943-4	GenBank	2	Complete	1581461	38.86	1503	92	6	37	1	EpsiloFsa, Epsilo Faa	Heli

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	conten	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
H. pylori	F57	AP011945	GenBank	1	Complete	1609006	38.73	1527	95	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	FD423	AKHM02000000	GenBank	115	Draft	1623128	39.15	1519	86	8	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	FD430	AKHN02000000	GenBank	129	Draft	1640355	39.1	1545	88	8	35	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	FD506	AKHO02000000	GenBank	105	Draft	1615241	38.74	1514	89	7	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	FD535	AKHP02000000	GenBank	81	Draft	1671564	39.08	1563	96	7	39	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	FD568	AKHQ02000000	GenBank	113	Draft	1610163	38.71	1502	87	8	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	FD577	AKHR02000000	GenBank	74	Draft	1625905	38.67	1540	88	7	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	FD662	AKHT02000000	GenBank	66	Draft	1659453	38.94	1554	95	5	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	FD703	AKHS02000000	GenBank	88	Draft	1676483	39.03	1555	89	7	45	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	FD719	AKHU02000000	GenBank	79	Draft	1642171	39.1	1554	96	7	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	G27	CP001173-4	GenBank	2	Complete	1663013	38.87	1574	95	7	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM100Ai	ANFP01000000	GenBank	80	Draft	1637801	39.29	1521	100	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM101Biv	APCW01000000	GenBank	78	Draft	1615843	39.29	1508	87	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM103Bi	APCX01000000	GenBank	71	Draft	1624174	39.27	1496	98	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM105Ai	APCY01000000	GenBank	82	Draft	1671331	39.22	1545	98	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM112Ai	APCZ01000000	GenBank	136	Draft	1625200	39.3	1494	92	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM114Ai	APDA01000000	GenBank	64	Draft	1622655	39.35	1494	96	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM115Ai	APDB01000000	GenBank	90	Draft	1684788	39.13	1565	96	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM117Ai	AWER01000000	GenBank	98	Draft	1643967	39.22	1514	95	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM118Bi	APDD01000000	GenBank	95	Draft	1683430	39.1	1545	99	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM119Bi	APDE01000000	GenBank	78	Draft	1661976	39.2	1542	99	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM120Ai	APDF01000000	GenBank	75	Draft	1687633	39.14	1557	101	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM121Aii	APDG01000000	GenBank	81	Draft	1662260	39.2	1520	104	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM201Ai	APDC01000000	GenBank	77	Draft	1615362	39.37	1481	89	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM210Bi	APDH01000000	GenBank	91	Draft	1621201	39.27	1499	87	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM231Ai	APDI01000000	GenBank	98	Draft	1624899	39.25	1497	90	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM239Bi	APDJ01000000	GenBank	91	Draft	1634248	39.2	1520	88	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM244Ai	APDK01000000	GenBank	95	Draft	1598086	39.35	1498	88	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM245Ai	APDL01000000	GenBank	78	Draft	1653356	39.16	1521	103	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM246Ai	APDM01000000	GenBank	102	Draft	1668715	39.18	1540	94	2	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM249T	APDN01000000	GenBank	82	Draft	1628240	39.26	1508	96	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM250AFi	APDO01000000	GenBank	71	Draft	1584388	39.39	1467	95	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM250T	APDP01000000	GenBank	73	Draft	1584136	39.39	1466	92	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM252Bi	APDQ01000000	GenBank	72	Draft	1582615	39.39	1465	95	3	37	1	EpsiloFsa, Epsilo Faa	Heli

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	conten	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
H. pylori	GAM252T	APDR01000000	GenBank	78	Draft	1583280	39.39	1466	93	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM254Ai	APDS01000000	GenBank	75	Draft	1641633	39.26	1520	98	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM260ASi	APDT01000000	GenBank	86	Draft	1633786	39.27	1513	99	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM260Bi	APDU01000000	GenBank	87	Draft	1671355	39.18	1547	100	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM260BSi	APDV01000000	GenBank	89	Draft	1578652	39.45	1478	83	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM263BFi	APDW01000000	GenBank	75	Draft	1652747	39.17	1518	97	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM264Ai	APDX01000000	GenBank	92	Draft	1607852	39.4	1489	87	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM265BSii	APDY01000000	GenBank	91	Draft	1664801	39.26	1536	90	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM268Bii	APDZ01000000	GenBank	124	Draft	1634314	39.27	1518	97	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM270ASi	APEA01000000	GenBank	87	Draft	1648901	39.21	1522	95	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM42Ai	APEB01000000	GenBank	80	Draft	1623480	39.31	1494	93	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM71Ai	APEC01000000	GenBank	124	Draft	1604806	39.3	1476	86	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM80Ai	APED01000000	GenBank	74	Draft	1641994	39.2	1501	98	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM83Bi	APEE01000000	GenBank	81	Draft	1611559	39.4	1514	85	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM83T	APEF01000000	GenBank	83	Draft	1613100	39.4	1515	84	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM93Bi	APEG01000000	GenBank	136	Draft	1622284	39.28	1497	87	2	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAM96Ai	APEH01000000	GenBank	112	Draft	1661367	39.38	1535	98	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Gambia94/24	CP002332-3	GenBank	2	Complete	1712468	39.12	1586	106	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAMchJs106B	APEI01000000	GenBank	61	Draft	1579819	39.45	1453	92	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAMchJs114i	APEJ01000000	GenBank	76	Draft	1616868	39.3	1499	95	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAMchJs117Ai	APEK01000000	GenBank	73	Draft	1617104	39.27	1478	95	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAMchJs124i	APEL01000000	GenBank	76	Draft	1611701	39.28	1483	95	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GAMchJs136i	APEM01000000	GenBank	86	Draft	1660131	39.25	1540	98	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	GC26	AKHV02000000	GenBank	111	Draft	1622472	38.73	1533	94	8	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	H13-1	AYUH01000000	GenBank	96	Draft	1631474	38.67	1869	92	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	H3014	CCMU01000000	GenBank	51	Draft	1642266	38.85	1545	91	4	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	H3016	CCMT01000000	GenBank	101	Draft	1642760	38.89	1638	83	4	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	H3017	CCMV01000000	GenBank	48	Draft	1650195	38.82	1748	86	4	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	H3018	CCMW01000000	GenBank	50	Draft	1655748	38.79	1559	95	4	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HLJ039	JAAA01000000	GenBank	34	Draft	1609997	38.71	1551	91	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HLJHP193	ALJI01000000	GenBank	55	Draft	1552322	38.91	1481	87	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HLJHP253	ALKC01000000	GenBank	20	Draft	1587316	38.82	1515	82	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HLJHP256	ALKA01000000	GenBank	40	Draft	1561171	38.92	1487	94	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HLJHP271	ALKB01000000	GenBank	43	Draft	1571832	38.82	1513	82	2	37	1	EpsiloFsa, Epsilo Faa	Heli

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	conten	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
H. pylori	Hp A-11	AOTW01000000	GenBank	1	Complete	1668342	38.84	1587	103	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp A-14	AKOT01000000	GenBank	8	Draft	1599343	39.06	1506	85	7	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp A-16	AKOU01000000	GenBank	11	Draft	1637800	39.19	1536	96	5	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp A-17	AKOD01000000	GenBank	6	Draft	1641046	39.24	1534	98	6	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp A-20	AKOE01000000	GenBank	10	Draft	1674600	39.21	1588	100	9	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp A-26	AKOV01000000	GenBank	9	Draft	1624194	38.97	1549	90	5	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp A-27	AKOW01000000	GenBank	8	Draft	1652101	38.84	1579	92	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp A-4	AKOA01000000	GenBank	9	Draft	1669747	39.27	1570	95	8	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp A-5	AKOB01000000	GenBank	7	Draft	1635925	39.23	1544	101	5	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp A-6	AKOR01000000	GenBank	12	Draft	1653378	39.22	1539	105	7	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp A-8	AKOS01000000	GenBank	7	Draft	1640628	39.22	1529	98	8	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp A-9	AKOC01000000	GenBank	18	Draft	1720407	38.73	1627	105	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-1	AOTX01000000	GenBank	54	Draft	1682934	39.01	1555	95	2	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-10	AKPB01000000	GenBank	32	Draft	1649818	39.24	1518	94	11	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-11	AKPC01000000	GenBank	10	Draft	1663168	38.96	1561	93	11	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-16	AKOF01000000	GenBank	9	Draft	1709635	39.03	1567	107	5	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-18	AKPD01000000	GenBank	21	Draft	1760691	38.99	1664	110	4	43	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-19	AKPE01000000	GenBank	7	Draft	1634366	39.31	1510	102	8	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-21	AKPF01000000	GenBank	7	Draft	1629534	39.21	1514	98	7	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-23	AKPG01000000	GenBank	9	Draft	1648064	39.14	1534	105	4	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-24	AKOG01000000	GenBank	8	Draft	1670182	39.17	1560	104	9	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-24b	AKPY01000000	GenBank	10	Draft	1668016	39.16	1557	102	8	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-24c	AKPZ01000000	GenBank	9	Draft	1670060	39.17	1555	104	9	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-27	AKOH01000000	GenBank	5	Draft	1612044	38.95	1525	86	5	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-28	AKOI01000000	GenBank	16	Draft	1625000	38.99	1534	87	9	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-29	AKOJ01000000	GenBank	16	Draft	1677607	39.12	1529	103	9	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-3	AKOX01000000	GenBank	11	Draft	1712042	38.93	1553	101	10	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-30	AKOK01000000	GenBank	13	Draft	1630837	39.39	1508	99	10	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-34	AKPH01000000	GenBank	8	Draft	1627237	39.23	1528	94	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-36	AKOL01000000	GenBank	16	Draft	1675907	39.08	1556	99	7	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-4	AKOY01000000	GenBank	9	Draft	1665093	39.13	1558	104	8	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-41	AKOM01000000	GenBank	11	Draft	1660567	39.22	1548	102	10	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-42	AKON01000000	GenBank	19	Draft	1702292	39.11	1567	104	8	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-43	AKOO01000000	GenBank	12	Draft	1607665	39.12	1500	86	7	36	1	EpsiloFsa, Epsilo Faa	Heli

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	conten	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
H. pylori	Hp H-44	AKOP01000000	GenBank	16	Draft	1665840	39.12	1536	105	5	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-45	AKOQ01000000	GenBank	14	Draft	1655397	38.95	1558	92	11	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-5b	AKQF01000000	GenBank	11	Draft	1708533	38.98	1585	105	2	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-6	AKOZ01000000	GenBank	10	Draft	1707985	38.96	1613	102	9	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp H-9	AKPA01000000	GenBank	12	Draft	1643101	38.87	1533	100	4	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp M1	AKQN01000000	GenBank	9	Draft	1660803	39.15	1558	103	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp M2	AKQO01000000	GenBank	11	Draft	1663213	39.15	1556	103	4	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp M3	AKQP01000000	GenBank	10	Draft	1676995	39.18	1564	102	10	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp M4	AKQQ01000000	GenBank	11	Draft	1671765	39.19	1557	104	8	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp M5	AKQR01000000	GenBank	7	Draft	1672403	39.18	1552	102	8	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp M6	AKQS01000000	GenBank	11	Draft	1669470	39.16	1565	104	6	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp M9	AKQT01000000	GenBank	11	Draft	1664471	39.17	1548	101	6	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-1	AKPI01000000	GenBank	6	Draft	1669637	39.09	1554	104	4	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-11	AKPN01000000	GenBank	12	Draft	1693878	39.08	1609	105	9	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-11b	AKQH01000000	GenBank	13	Draft	1695255	39.08	1586	106	10	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-13	AKPO01000000	GenBank	10	Draft	1714414	39.03	1661	101	9	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-13b	AKQI01000000	GenBank	8	Draft	1711532	39.01	1572	102	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-15	AKPP01000000	GenBank	4	Draft	1652821	38.91	1557	91	10	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-15b	AKQJ01000000	GenBank	7	Draft	1658647	38.93	1553	93	11	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-16	AKPQ01000000	GenBank	5	Draft	1549132	39.11	1466	84	4	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-1b	AKQA01000000	GenBank	8	Draft	1672266	39.12	1561	103	2	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-2	AKPJ01000000	GenBank	6	Draft	1693870	39.03	1565	106	5	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-23	AKPR01000000	GenBank	3	Draft	1643525	38.81	1580	90	4	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-25	AKPS01000000	GenBank	3	Draft	1667461	39.2	1690	98	5	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-25c	AKQK01000000	GenBank	6	Draft	1669110	39.18	1554	105	5	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-25d	AKQL01000000	GenBank	10	Draft	1662539	39.19	1547	104	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-26	AKPT01000000	GenBank	9	Draft	1697203	38.98	1584	105	5	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-28b	AKQM01000000	GenBank	12	Draft	1625539	39.25	1527	95	10	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-2b	AKQB01000000	GenBank	8	Draft	1714545	39.04	1588	106	5	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-3	AKPK01000000	GenBank	11	Draft	1656906	39.21	1549	103	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-30	AKPU01000000	GenBank	5	Draft	1640856	38.84	1539	94	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-3b	AKQC01000000	GenBank	9	Draft	1661739	39.23	1538	103	10	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-4	AKPL01000000	GenBank	9	Draft	1697737	39.14	1566	102	8	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-41	AKPV01000000	GenBank	10	Draft	1716127	39.13	1602	106	11	38	1	EpsiloFsa, Epsilo Faa	Heli

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	content	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
H. pylori	Hp P-4c	AKQD01000000	GenBank	8	Draft	1683310	39.11	1568	100	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-4d	AKQE01000000	GenBank	9	Draft	1682896	39.09	1560	100	4	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-62	AKPW01000000	GenBank	13	Draft	1653470	39.15	1528	101	7	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-74	AKPX01000000	GenBank	2	Draft	1622385	38.96	1520	93	5	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-8	AKPM01000000	GenBank	7	Draft	1615626	39.24	1508	92	5	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Hp P-8b	AKQG01000000	GenBank	6	Draft	1625223	39.26	1513	95	5	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HP116Bi	APEN01000000	GenBank	72	Draft	1652421	39.23	1520	95	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HP250AFii	APEO01000000	GenBank	76	Draft	1584486	39.39	1466	92	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HP250AFiii	APEP01000000	GenBank	74	Draft	1583357	39.4	1462	94	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HP250AFiv	APEQ01000000	GenBank	73	Draft	1585052	39.39	1468	94	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HP250ASi	APER01000000	GenBank	77	Draft	1583980	39.39	1465	94	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HP250ASii	APES01000000	GenBank	78	Draft	1584815	39.39	1467	94	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HP250BFi	APET01000000	GenBank	135	Draft	1591331	39.38	1472	93	2	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HP250BFii	APEU01000000	GenBank	73	Draft	1587569	39.39	1472	95	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HP250BFiii	APEV01000000	GenBank	77	Draft	1586191	39.39	1467	94	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HP250BFiV	APEW01000000	GenBank	91	Draft	1583907	39.39	1463	89	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HP250BSi	APEX01000000	GenBank	93	Draft	1582781	39.4	1461	90	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HP260AFi	APEY01000000	GenBank	75	Draft	1637940	39.28	1516	99	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HP260AFii	APEZ01000000	GenBank	81	Draft	1636368	39.28	1515	98	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HP260ASii	APFA01000000	GenBank	86	Draft	1638096	39.28	1515	97	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HP260BFii	APFB01000000	GenBank	72	Draft	1581856	39.45	1482	83	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HP260Bi	APFC01000000	GenBank	82	Draft	1672596	39.18	1547	101	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HP87hu	CBRI01000000	GenBank	72	Draft	1637756	38.9	1573	92	4	38	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HP87P7	CBRJ01000000	GenBank	46	Draft	1664575	38.72	1594	95	4	37	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	HP87P7tlpDRI	CBRL01000000	GenBank	46	Draft	1664988	38.72	1579	97	4	37	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	HP87tlpD	CBRK01000000	GenBank	49	Draft	1664475	38.72	1589	97	4	37	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	HPAG1	CP000241-2	GenBank	2	Complete	1605736	39.07	1505	104	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HPARG63	CBKY01000000	GenBank	34	Draft	1668677	38.8	1553	93	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HPARG8G	CBKZ01000000	GenBank	47	Draft	1602229	39.04	1509	96	3	35	2	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HPKX_438_AG0C1	ABJO01000000	GenBank	2602	Draft	1819663	39.49	2512	53	4	36	0	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HPKX_438_CA4C1	ABJP01000000	GenBank	3766	Draft	1566388	39.19	3858	32	3	32	0	EpsiloFsa, Epsilo Faa	Heli
H. pylori	HUP-B14	CP003486-7	GenBank	2	Complete	1607584	39.04	1506	97	6	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	India7	CP002331	GenBank	1	Complete	1675918	38.9	1568	92	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	J166	CP007603	GenBank	1	Complete	1650561	38.93	1538	101	6	37	1	EpsiloFsa, Epsilo Faa	Heli

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	conten	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
H. pylori	J166output_1moA	JMGE01000000	GenBank	43	Draft	1632538	38.93	1519	97	3	37	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	J166output_1moB	JMGF01000000	GenBank	46	Draft	1632050	38.93	1524	98	3	37	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	J166output_1moC	JMGG01000000	GenBank	49	Draft	1632236	38.92	1523	97	3	37	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	J166output_1wkA	JMGN01000000	GenBank	49	Draft	1634015	38.94	1522	98	3	37	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	J166output_1wkB	JMGC01000000	GenBank	47	Draft	1634094	38.94	1523	98	3	37	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	J166output_1wkC	JMGD01000000	GenBank	47	Draft	1634547	38.94	1522	98	3	37	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	J166output_2moA	JMGH01000000	GenBank	41	Draft	1630637	38.93	1517	97	3	37	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	J166output_2moB	JMGI01000000	GenBank	47	Draft	1630429	38.93	1518	96	3	37	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	J166output_2moC	JMGJ01000000	GenBank	46	Draft	1634811	38.94	1521	97	3	37	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	J166output_6moA	JMGK01000000	GenBank	50	Draft	1634018	38.94	1521	94	3	37	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	J166output_6moB	JMGL01000000	GenBank	56	Draft	1640402	38.96	1521	94	3	37	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	J166output_6moC	JMGM01000000	GenBank	51	Draft	1634374	38.94	1524	92	3	37	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	J99	AE001439	GenBank	1	Complete	1643831	39.19	1504	101	6	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Lithuania75	CP002334-5	GenBank	2	Complete	1640673	38.87	1550	92	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Manado-1	JRAC01000000	GenBank	38	Draft	1539815	38.85	1477	81	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	N6	CAHX01000000	GenBank	54	Draft	1657684	38.74	1638	97	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NAB47	AJFA02000000	GenBank	107	Draft	1585921	39.14	1473	88	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NAD1	AJGJ02000000	GenBank	103	Draft	1603561	38.92	1504	91	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NAK7	AONJ01000000	GenBank	68	Draft	1586710	39.04	1481	86	4	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NCTC 11637^{T}	AIHX01000000	GenBank	163	Draft	1600513	39	1526	83	2	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NQ1671	CADM01000000	GenBank	38	Draft	1626950	39.08	1516	92	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NQ1701	CADH01000000	GenBank	78	Draft	1639623	38.95	1802	85	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NQ1707	CADJ01000000	GenBank	91	Draft	1654738	38.9	1521	95	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NQ1712	CADF01000000	GenBank	62	Draft	1572582	39.08	1495	78	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NQ315	CADE01000000	GenBank	57	Draft	1601938	39.02	1517	85	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NQ352	CADG01000000	GenBank	61	Draft	1639151	38.95	1600	90	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NQ367	CADL01000000	GenBank	90	Draft	1622411	39.09	1601	88	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NQ392	CADI01000000	GenBank	80	Draft	1654988	38.9	1546	97	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NQ4044	AKNW01000000	GenBank	17	Draft	1728524	38.7	1648	103	5	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NQ4053	AKNV01000000	GenBank	6	Draft	1652030	38.89	1557	94	6	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NQ4060	CADK01000000	GenBank	59	Draft	1649736	38.9	1547	94	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NQ4076	AKNX01000000	GenBank	4	Draft	1632709	38.98	1526	88	7	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NQ4099	AKNU01000000	GenBank	6	Draft	1650644	39.01	1531	89	9	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NQ4110	AKNZ01000000	GenBank	3	Draft	1602302	39.17	1501	85	6	36	1	EpsiloFsa, Epsilo Faa	Heli

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	conten	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
H. pylori	NQ4161	AKNY01000000	GenBank	9	Draft	1643218	39.01	1565	91	9	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NQ4191	CADN01000000	GenBank	43	Draft	1629065	39.06	1524	92	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NQ4200	AKNS01000000	GenBank	14	Draft	1646737	38.97	1524	98	6	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NQ4216	AKNR01000000	GenBank	13	Draft	1655890	38.94	1553	95	5	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NQ4228	AKNT01000000	GenBank	6	Draft	1653281	38.96	1527	105	4	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	NY40	AP014523	GenBank	1	Complete	1696917	38.81	1626	99	6	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	OK113	AP012600	GenBank	1	Complete	1616617	38.73	1536	92	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	OK310	AP012601-2	GenBank	2	Complete	1595436	38.77	1511	91	6	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	oki102	CP006820	GenBank	1	Complete	1633212	38.81	1524	95	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	oki112	CP006821	GenBank	1	Complete	1637925	38.81	1532	101	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	oki128	CP006822	GenBank	1	Complete	1553826	38.97	1492	77	6	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	oki154	CP006823	GenBank	1	Complete	1599700	38.8	1532	79	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	oki422	CP006824	GenBank	1	Complete	1634852	38.83	1534	95	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	oki673	CP006825	GenBank	1	Complete	1595058	38.82	1529	75	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	oki828	CP006826	GenBank	1	Complete	1600345	38.8	1536	83	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	oki898	CP006827	GenBank	1	Complete	1634875	38.83	1513	94	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	P12	CP001217-8	GenBank	2	Complete	1684038	38.79	1584	101	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	P79	AIHW01000000	GenBank	215	Draft	1617653	38.89	1547	78	2	35	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	PeCan18	CP003475	GenBank	1	Complete	1660685	39.02	1540	97	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	PeCan4	CP002074-5	GenBank	2	Complete	1638269	38.91	1529	90	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Puno120	CP002980-1	GenBank	2	Complete	1637762	38.9	1529	87	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Puno135	CP002982	GenBank	1	Complete	1646139	38.82	1536	87	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	PZ5004	ASZF01000000	GenBank	259	Draft	1586499	38.76	1533	78	2	32	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	PZ5024	ASYS01000000	GenBank	360	Draft	1522037	38.33	1519	72	1	25	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	PZ5026	ASYT01000000	GenBank	224	Draft	1622716	38.74	1528	87	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	PZ5056	ASYU01000000	GenBank	298	Draft	1603341	38.72	1572	87	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	PZ5080	ASYV01000000	GenBank	244	Draft	1613777	38.63	1556	86	2	34	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	PZ5086	ASYW01000000	GenBank	244	Draft	1559914	38.78	1493	73	2	33	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	R018c	AMOQ01000000	GenBank	15	Draft	1646836	38.87	1555	99	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	R030b	AMOR01000000	GenBank	11	Draft	1610617	39.21	1510	97	9	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	R036d	AMOT01000000	GenBank	8	Draft	1640262	38.92	1541	97	8	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	R037c	AMOU01000000	GenBank	9	Draft	1614130	38.9	1527	81	8	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	R038b	AMOV01000000	GenBank	14	Draft	1632757	38.96	1521	90	8	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	R046Wa	AMOW01000000	GenBank	8	Draft	1585443	39.08	1498	90	9	37	1	EpsiloFsa, Epsilo Faa	Heli

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	content	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
H. pylori	R055a	AMOX01000000	GenBank	16	Draft	1631090	38.87	1540	100	8	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	R056a	AMOY01000000	GenBank	13	Draft	1646110	38.88	1551	101	5	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	R32b	AMOS01000000	GenBank	10	Draft	1582461	38.94	1502	92	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Rif1	CP003905	GenBank	1	Complete	1667883	38.87	1574	95	7	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Rif2	CP003906	GenBank	1	Complete	1667890	38.87	1574	95	7	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA144A	CBNJ01000000	GenBank	39	Draft	1630166	38.55	1615	78	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA144C	CBOX01000000	GenBank	30	Draft	1628642	38.54	1610	77	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA146A	CBPZ01000000	GenBank	45	Draft	1673450	38.89	1569	104	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA146C	CBPF01000000	GenBank	46	Draft	1672384	38.9	1704	101	4	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA155A	CBNZ01000000	GenBank	33	Draft	1664256	38.4	1651	85	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA155C	CBOH01000000	GenBank	57	Draft	1664590	38.41	1562	84	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA156A	CBNR01000000	GenBank	50	Draft	1699733	38.96	1611	95	4	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA156C	CBPV01000000	GenBank	75	Draft	1698931	38.97	1583	93	4	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA157A	CBNF01000000	GenBank	72	Draft	1670410	38.93	1561	95	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA157C	CBPJ01000000	GenBank	78	Draft	1675411	38.91	1857	88	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA158A	CBOL01000000	GenBank	36	Draft	1614685	39.09	1567	92	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA158C	CBPY01000000	GenBank	55	Draft	1614176	39.1	1490	94	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA160A	CBQN01000000	GenBank	50	Draft	1658851	38.51	1535	84	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA160C	CBQF01000000	GenBank	50	Draft	1657073	38.52	1571	83	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA161A	CBPE01000000	GenBank	49	Draft	1653991	39.03	1597	97	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA161C	CBNM01000000	GenBank	62	Draft	1689135	38.97	1584	101	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA162A	CBOS01000000	GenBank	71	Draft	1667382	38.98	1558	90	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA162C	CBNQ01000000	GenBank	103	Draft	1647112	39.09	1548	92	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA163A	CBOG01000000	GenBank	48	Draft	1616449	39.09	1783	90	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA163C	CBOW01000000	GenBank	35	Draft	1617889	39.09	1503	94	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA164A	CBQK01000000	GenBank	36	Draft	1616748	38.83	1583	82	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA164C	CBQM01000000	GenBank	55	Draft	1613002	38.85	1506	80	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA165A	CBQD01000000	GenBank	36	Draft	1616887	38.83	1531	82	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA165C	CBQJ01000000	GenBank	81	Draft	1631585	38.84	1546	78	3	40	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA166A	CBPA01000000	GenBank	48	Draft	1599095	38.6	1548	78	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA168A	CBPU01000000	GenBank	74	Draft	1642196	39	1536	95	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA168C	CBOC01000000	GenBank	71	Draft	1646957	38.97	1559	94	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA169A	CBQH01000000	GenBank	125	Draft	1652206	38.48	1589	83	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA169C	CBOK01000000	GenBank	55	Draft	1661612	38.4	1575	84	3	36	1	EpsiloFsa, Epsilo Faa	Heli

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	conten	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
H. pylori	SA170A	CBMZ01000000	GenBank	64	Draft	1607873	39.24	1514	87	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA170C	CBQL01000000	GenBank	53	Draft	1611872	39.21	1507	89	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA171A	CBOA01000000	GenBank	36	Draft	1614279	38.84	1502	86	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA171C	CBPC01000000	GenBank	37	Draft	1611340	38.83	1501	85	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA172A	CBOY01000000	GenBank	32	Draft	1613609	38.83	1502	86	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA172C	CBPG01000000	GenBank	51	Draft	1658069	38.44	1528	87	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA173A	CBOU01000000	GenBank	82	Draft	1647792	38.9	1524	83	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA173C	CBNG01000000	GenBank	59	Draft	1657037	38.83	1545	84	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA174A	CBPS01000000	GenBank	91	Draft	1652635	38.49	1545	85	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA175A	CBPO01000000	GenBank	40	Draft	1612413	38.54	1489	80	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA175C	CBOI01000000	GenBank	26	Draft	1617911	38.54	1505	80	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA194A	CBPW01000000	GenBank	46	Draft	1565205	38.7	1458	82	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA194C	CBNK01000000	GenBank	54	Draft	1566574	38.69	1470	83	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA210A	CBNW01000000	GenBank	45	Draft	1658916	38.94	1865	91	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA210C	CBOQ01000000	GenBank	46	Draft	1662295	38.93	1533	98	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA213A	CBOM01000000	GenBank	86	Draft	1625887	38.78	1500	89	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA213C	CBPK01000000	GenBank	100	Draft	1618554	38.84	1499	91	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA214A	CBOE01000000	GenBank	55	Draft	1631197	39.06	1523	94	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA214C	CBNS01000000	GenBank	70	Draft	1652641	39.05	1548	97	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA215C	CBNO01000000	GenBank	68	Draft	1665140	39.02	1546	99	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA216C	CBNC01000000	GenBank	160	Draft	1634140	39.18	1614	90	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA220A	CBND01000000	GenBank	78	Draft	1664601	38.98	1540	98	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA220C	CBPX01000000	GenBank	65	Draft	1666482	38.96	1540	97	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA221A	CBQA01000000	GenBank	53	Draft	1624426	38.86	1516	85	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA221C	CBPT01000000	GenBank	42	Draft	1625684	38.85	1516	86	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA222A	CBQE01000000	GenBank	62	Draft	1658965	38.92	1612	89	4	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA222C	CBON01000000	GenBank	38	Draft	1630022	38.95	1553	82	4	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA224A	CBOB01000000	GenBank	55	Draft	1645150	38.98	1799	90	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA224C	CBPP01000000	GenBank	62	Draft	1648350	38.98	1851	85	2	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA226A	CBNH01000000	GenBank	59	Draft	1705370	38.84	1607	97	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA227A	CBNX01000000	GenBank	65	Draft	1662052	39.04	1679	92	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA227C	CBNL01000000	GenBank	64	Draft	1664983	39.06	1553	99	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA233A	CBMX01000000	GenBank	49	Draft	1578535	38.69	1473	85	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA233C	CBPD01000000	GenBank	43	Draft	1581472	38.69	1471	83	3	36	1	EpsiloFsa, Epsilo Faa	Heli

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	content	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
H. pylori	SA251A	CBQC01000000	GenBank	35	Draft	1635579	38.58	1536	81	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA251C	CBOZ01000000	GenBank	53	Draft	1631565	38.6	1515	83	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA252A	CBNP01000000	GenBank	59	Draft	1636149	38.99	1672	91	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA252C	CBOV01000000	GenBank	54	Draft	1652831	38.99	1708	95	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA253A	CBOJ01000000	GenBank	57	Draft	1647488	38.46	1669	79	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA253C	CBOR01000000	GenBank	44	Draft	1646187	38.45	1530	85	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA29A	CBMV01000000	GenBank	42	Draft	1620792	38.6	1696	72	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA29C	CBNI01000000	GenBank	93	Draft	1650643	38.98	1813	83	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA300A	CBQG01000000	GenBank	40	Draft	1615195	39.1	1670	90	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA300C	CBPL01000000	GenBank	36	Draft	1617352	39.09	1497	94	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA301A	CBPH01000000	GenBank	51	Draft	1657659	39.03	1570	101	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA301C	CBNT01000000	GenBank	102	Draft	1641456	39.14	1544	102	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA302A	CBOF01000000	GenBank	63	Draft	1563532	39.05	1478	82	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA302C	CBNA01000000	GenBank	49	Draft	1571928	39	1474	83	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA303C	CBMW01000000	GenBank	88	Draft	1638342	38.6	1517	86	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA30A	CBPQ01000000	GenBank	64	Draft	1670386	38.89	1585	96	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA30C	CBQB01000000	GenBank	94	Draft	1664923	38.91	1569	94	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA31C	CBPM01000000	GenBank	47	Draft	1615181	39.09	1491	94	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA34A	CBPI01000000	GenBank	31	Draft	1644852	38.46	1532	92	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA34C	CBNY01000000	GenBank	131	Draft	1548190	38.89	1430	80	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA35A	CBNE01000000	GenBank	64	Draft	1659464	38.88	1556	90	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA35C	CBNU01000000	GenBank	43	Draft	1659115	38.89	1559	93	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA36C	CBQI01000000	GenBank	57	Draft	1640611	38.47	1535	81	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA37A	CBOO01000000	GenBank	46	Draft	1600489	38.95	1496	86	4	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA37C	CBOP01000000	GenBank	106	Draft	1545458	39.23	1437	84	4	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA40A	CBNN01000000	GenBank	48	Draft	1646000	38.46	1537	89	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA40C	CBPN01000000	GenBank	188	Draft	1528413	38.97	1406	77	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA45A	CBNB01000000	GenBank	56	Draft	1638532	39.01	1525	96	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA45C	CBPR01000000	GenBank	47	Draft	1640985	39	1524	96	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA46A	CBNV01000000	GenBank	70	Draft	1599857	39.16	1482	94	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA46C	CBOD01000000	GenBank	47	Draft	1607590	39.13	1637	92	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA47A	CBOT01000000	GenBank	43	Draft	1670767	38.43	1576	85	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SA47C	CBPB01000000	GenBank	29	Draft	1669624	38.44	1782	87	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Sahul64	ALWV01000000	GenBank	54	Draft	1644275	38.76	1546	82	3	37	1	EpsiloFsa, Epsilo Faa	Heli

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	conten	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
H. pylori	Sat464	CP002071-2	GenBank	2	Complete	1567570	39.09	1461	86	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Shi112	CP003474	GenBank	1	Complete	1663456	38.77	1558	96	6	38	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Shi169	CP003473	GenBank	1	Complete	1616909	38.86	1514	89	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Shi417	CP003472	GenBank	1	Complete	1665719	38.77	1541	93	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Shi470	CP001072	GenBank	1	Complete	1608548	38.91	1517	95	6	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SJM180	CP002073	GenBank	1	Complete	1658051	38.9	1520	100	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SNT49	CP002983-4	GenBank	2	Complete	1610830	39	1511	90	6	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SouthAfrica20	CP006691	GenBank	1	Complete	1622867	38.57	1579	88	6	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SouthAfrica50	AVNI01000000	GenBank	2	Draft	1600214	38.62	1763	82	6	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	SouthAfrica7	CP002336-7	GenBank	2	Complete	1679829	38.42	1569	94	6	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	Taiwan-47	JQNY01000000	GenBank	58	Draft	1577473	38.73	1514	86	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	UM007	AONM01000000	GenBank	72	Draft	1575854	38.81	1489	90	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	UM018	AONK01000000	GenBank	72	Draft	1614546	39.06	1513	88	4	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	UM023	AUSK01000000	GenBank	34	Draft	1623075	38.76	1531	89	10	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	UM032	CP005490	GenBank	1	Complete	1593537	38.82	1511	94	6	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	UM034	AONN01000000	GenBank	65	Draft	1608430	38.72	1496	86	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	UM037	AUSI01000000	GenBank	60	Draft	1723664	38.92	1591	91	12	36	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	UM037	CP005492	GenBank	1	Complete	1692794	38.89	1595	93	6	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	UM038	AUSL01000000	GenBank	44	Draft	1762049	38.43	1615	93	8	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	UM045	AONO01000000	GenBank	59	Draft	1594207	39.01	1485	91	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	UM054	AONL01000000	GenBank	81	Draft	1594474	39.11	1483	88	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	UM065	AUSM01000000	GenBank	38	Draft	1586653	38.9	1469	88	10	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	UM066	AUSJ01000000	GenBank	34	Draft	1694163	38.64	1586	97	12	37	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	UM066	CP005493	GenBank	1	Complete	1658047	38.62	1573	98	6	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	UM067	AUSN01000000	GenBank	44	Draft	1680838	39.04	1568	102	13	38	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	UM077	AUSQ01000000	GenBank	53	Draft	1619377	38.82	1519	96	12	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	UM084	AUSO01000000	GenBank	34	Draft	1656826	39.05	1535	88	12	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	UM085	AUSP01000000	GenBank	50	Draft	1645062	38.73	1522	102	12	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	UM111	AUSR01000000	GenBank	38	Draft	1663127	38.68	1553	99	12	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	UM114	AUSS01000000	GenBank	36	Draft	1708937	38.92	1586	97	11	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	UM298	CP006610	GenBank	1	Complete	1594544	38.82	1515	94	6	36	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	UM299	CP005491	GenBank	1	Complete	1594569	38.82	1515	94	6	36	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. pylori	UMB_G1	AOTV01000000	GenBank	5	Draft	1571801	39.24	1469	89	8	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	v225d	CP001582-3	GenBank	2	Complete	1595604	38.94	1509	86	6	36	1	EpsiloFsa, Epsilo Faa	Heli

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	content	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
H. pylori	wls-5-1	JCKE01000000	GenBank	72	Draft	1632482	38.7	1541	86	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	wls-5-10	JDVP01000000	GenBank	71	Draft	1632794	38.68	1532	90	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	wls-5-11	JDVO01000000	GenBank	71	Draft	1569616	38.83	1485	88	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	wls-5-12	AUPY01000000	GenBank	65	Draft	1582646	38.79	1505	87	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	wls-5-13	JDVN01000000	GenBank	35	Draft	1573853	38.8	1496	86	4	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	wls-5-14	JDVM01000000	GenBank	67	Draft	1567825	38.82	1489	90	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	wls-5-15	JDVL01000000	GenBank	78	Draft	1608460	38.76	1504	83	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	wls-5-16	JDVK01000000	GenBank	124	Draft	1602632	38.86	1506	83	3	35	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	wls-5-17	JDVJ01000000	GenBank	40	Draft	1564207	38.83	1489	86	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	wls-5-18	JDVI01000000	GenBank	48	Draft	1626553	38.68	1531	87	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	wls-5-2	JCKD01000000	GenBank	75	Draft	1621628	38.72	1516	85	3	35	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	wls-5-3	AUPD01000000	GenBank	91	Draft	1653877	38.68	1549	88	3	35	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	wls-5-4	JDVV01000000	GenBank	80	Draft	1611963	38.79	1513	87	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	wls-5-5	JDVU01000000	GenBank	88	Draft	1565074	38.82	1477	86	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	wls-5-6	JDVT01000000	GenBank	154	Draft	1535407	38.82	1462	76	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	wls-5-7	JDVS01000000	GenBank	255	Draft	1634324	38.89	1512	76	2	35	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	wls-5-8	JDVR01000000	GenBank	98	Draft	1618299	38.72	1531	84	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	wls-5-9	JDVQ01000000	GenBank	78	Draft	1629878	38.67	1526	87	3	34	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	X47-2AL	AWNG01000000	GenBank	63	Draft	1617514	38.92	1526	91	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	XZ274	CP003419-20	GenBank	2	Complete	1656544	38.57	1698	89	7	36	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	YN1-91	JPXC01000000	GenBank	11	Draft	1597629	38.7	1533	92	3	37	1	EpsiloFsa, Epsilo Faa	Heli
H. pylori	YN4-84	JPXD01000000	GenBank	9	Draft	1622416	38.69	1559	86	2	36	1	EpsiloFsa, Epsilo Faa	Heli
H. rodentium	ATCC 700285^T	JHWC01000000	GenBank	29	Draft	1810652	37.04	1791	100	3	37	1	EpsiloFsa, Epsilo Faa	Heli
"H. sanguini"	MIT 97-6194	JRMP01000000	GenBank	178	Draft	2920052	34.66	2638	159	4	39	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
H. suis	HS1^T	ADGY01000000	GenBank	136	Draft	1635292	39.91	1649	70	5	39	1	EpsiloFsa, Epsilo Faa	Heli
H. suis	HS5	ADHO01000000	GenBank	319	Draft	1669960	39.91	1672	69	5	39	1	EpsiloFsa, Epsilo Faa	Heli
H. trogontum	ATCC 49310	JRPK01000000	GenBank	657	Draft	2548826	33.23	2290	125	3	36	1	EpsiloFsa, Epsilo Faa	Heli
H. trogontum	ATCC 700114^T	JRPL01000000	GenBank	129	Draft	2765848	33.05	2549	167	6	38	1	EpsiloFsa, Epsilo Faa	Heli
H. typhlonius	MIT 98-6810^T	JRPF01000000	GenBank	25	Draft	1900025	38.86	1934	106	9	40	1	EpsiloFsa, Epsilo Faa	Heli
"H. winghamensis"	ATCC BAA-430	ACDO01000000	GenBank	55	Draft	1654865	35.49	1657	86	3	38	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
Helicobacter sp.	MIT 01-6451	JRMQ01000000	GenBank	48	Draft	2051807	35.76	2037	127	9	43	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
Helicobacter sp.	MIT 03-1614	JRMS01000000	GenBank	172	Draft	1946884	35.78	1934	108	3	39	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
Helicobacter sp.	MIT 03-1616	JROY01000000	GenBank	176	Draft	1910335	38.81	1863	100	5	38	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
Helicobacter sp.	MIT 05-5293	JROZ01000000	GenBank	101	Draft	2008733	37.66	2047	103	3	44	1	EpsiloFsa, Epsilo Faa	Heli (ignored)

Supplementary File 2

Taxon	Strain	Genome Name	Source of Sequence	Seque	Genome	Length (nt)	conten	No. CDS	Signal peptides	rRNA	tRNA	tmRNA	Local BLAST database	LS-BSR
Helicobacter sp.	MIT 05-5294	JRPA01000000	GenBank	3237	Draft	3219750	37.92	4250	92	2	36	1	EpsiloFsa, Epsilo Faa	Heli (ignored)
Helicobacter sp.	MIT 11-5569	JRPB01000000	GenBank	83	Draft	2023773	34.58	2085	120	9	42	2	EpsiloFsa, Epsilo Faa	Heli (ignored)
Lebetimonas sp.	JH292	ATHQ01000000	GenBank	5	Draft	1641000	31.43	1896	40	6	42	1	EpsiloFsa, Epsilo Faa	
Lebetimonas sp.	JH369	ATHP01000000	GenBank	5	Draft	1740444	31.33	1929	39	6	41	1	EpsiloFsa, Epsilo Faa	
Lebetimonas sp.	JS032	ATHS01000000	GenBank	5	Draft	1709134	31.49	1853	43	6	41	1	EpsiloFsa, Epsilo Faa	
Lebetimonas sp.	JS085	ATHR01000000	GenBank	5	Draft	1740329	31.33	1869	42	6	41	1	EpsiloFsa, Epsilo Faa	
Lebetimonas sp.	JS138	ATHT01000000	GenBank	5	Draft	1688674	31.51	1823	42	6	41	1	EpsiloFsa, Epsilo Faa	
Lebetimonas sp.	JS170	ATHU01000000	GenBank	5	Draft	1740475	31.33	1853	43	6	41	1	EpsiloFsa, Epsilo Faa	
Nautilia profundicola	AmH^T	CP001279	GenBank	1	Complete	1676444	33.51	1734	71	12	49	1	EpsiloFsa, Epsilo Faa	
Nitratifractor salsuginis	DSM 16511^T	CP002452	GenBank	1	Complete	2101285	53.91	2106	130	6	45	1	EpsiloFsa, Epsilo Faa	
Nitratiruptor sp.	SB155-2	AP009178	GenBank	1	Complete	1877931	39.69	1930	93	9	45	1	EpsiloFsa, Epsilo Faa	
Sulfuricurvum kujiense	DSM 16994^T	CP002355-9	GenBank	5	Complete	2819357	44.56	2815	201	9	49	1	EpsiloFsa, Epsilo Faa	
Sulfuricurvum sp.	MLSB	JQGL01000000	GenBank	148	Draft	2106130	49.11	2223	146	2	41	1	EpsiloFsa, Epsilo Faa	
Sulfurimonas autotrophica	DSM 16294^T	CP002205	GenBank	1	Complete	2153198	35.24	2161	148	12	43	1	EpsiloFsa, Epsilo Faa	
Sulfurimonas denitrificans	DSM 1251^T	CP000153	GenBank	1	Complete	2201561	34.46	2162	134	12	45	1	EpsiloFsa, Epsilo Faa	
Sulfurimonas gotlandica	GD1^T	ABXD01000000	GenBank	26	Draft	2939048	33.59	2851	243	12	48	1	EpsiloFsa, Epsilo Faa	
Sulfurimonas gotlandica	GD1^T	AFRZ01000000	GenBank	1	Complete	2952682	33.6	2874	250	12	48	1	EpsiloFsa, Epsilo Faa	
Sulfurimonas sp.	AST-10	AUPZ01000000	GenBank	28	Draft	2302023	34.89	2241	156	3	40	1	EpsiloFsa, Epsilo Faa	
Sulfurospirillum arcachonense	DSM 9755^T	JFBL01000000	GenBank	42	Draft	2656374	30.39	2640	157	5	34	1	EpsiloFsa, Epsilo Faa	
Sulfurospirillum arsenophilum	NBRC 109478^T	BBQF01000000	GenBank	11	Draft	2629984	39.24	2593	219	3	40	1	EpsiloFsa, Epsilo Faa	
Sulfurospirillum barnesii	SES-3^T	CP003333	GenBank	1	Complete	2510109	38.82	2511	164	6	43	1	EpsiloFsa, Epsilo Faa	
Sulfurospirillum cavolei	NBRC 109482^T	BBQE01000000	GenBank	27	Draft	2833185	43.82	2839	211	6	47	1	EpsiloFsa, Epsilo Faa	
Sulfurospirillum deleyianum	DSM 6946^T	CP001816	GenBank	1	Complete	2306351	38.97	2284	158	9	44	1	EpsiloFsa, Epsilo Faa	
Sulfurospirillum multivorans	DSM 12446^T	CP007201	GenBank	1	Complete	3175729	40.9	3302	238	6	46	1	EpsiloFsa, Epsilo Faa	
Sulfurospirillum sp.	MES	JSEC01000000	GenBank	61	Draft	2651881	43.82	2637	198	1	37	1	EpsiloFsa, Epsilo Faa	
Sulfurospirillum sp.	SCADC	JQGK01000000	GenBank	38	Draft	2661535	41.63	2705	195	1	40	1	EpsiloFsa, Epsilo Faa	
Sulfurovum sp.	AR	AJLE01000000	GenBank	11	Draft	2125769	39.24	2149	180	0	41	1	EpsiloFsa, Epsilo Faa	
Sulfurovum sp.	NBC37-1	AP009179	GenBank	1	Complete	2562277	43.87	2553	262	9	44	1	EpsiloFsa, Epsilo Faa	
Sulfurovum sp.	SCGC AAA036-F05	AQWF01000000	GenBank	107	Draft	611024	33.63	608	17	5	15	0	EpsiloFsa, Epsilo Faa	
Thiovulum sp.	ES	AKKQ01000000	GenBank	206	Draft	2080587	33.01	1995	64	3	36	1	EpsiloFsa, Epsilo Faa	
Wolinella succinogenes	DSM 1740^T	BX571657.1-BX571663.1	GenBank	7	Draft	2110655	48.46	2074	116	9	41	1	EpsiloFsa, Epsilo Faa	

Supplementary File 3

Sample Number	Date of Birth	Date of Collection	Case (1) or Control (0)	Matched Set	Gender (1=male)	Age (days)	C. jejuni	C. concisus	C. coli	Campylobactereaceae by Culture	Rotavirus	Other Taxa	Faecal MLPA Result	FaecalMLPA Jejuni	FaecalMLPA Concisus
10817726	30/05/12	6/05/15	1	123	0	1071	1	0	0	C. jejuni	0		Negative	0	0
10818452	18/11/14	8/05/15	1	23	0	171	0	0	0		0		Negative	0	0
10819077	31/03/15	11/05/15	1	4	1	41	0	0	0		0	Cl. difficile	Negative	0	0
10820448	18/11/08	13/05/15	1	155	0	2367	0	0	0		0	Blastocystis hominis, Dientamoeba	Negative	0	0
10825751	20/12/13	1/06/15	0	76	0	528	0	0	0		0		Not undertaken	NA	NA
10826643	30/04/14	2/06/15	1	62	1	398	0	0	0		1	Rorovirus	Not undertaken	NA	NA
10837941	22/09/13	25/06/15	1	93	0	641	0	0	0		0		Not undertaken	NA	NA
10838427	7/08/14	27/06/15	1	49	0	324	0	0	0		0	Adenovirus	Not undertaken	NA	NA
10841829	16/01/15	8/07/15	1	24	1	173	1	0	0	C. jejuni	0		C. jejuni	1	0
10843131	20/01/14	12/07/15	1	83	0	538	1	0	0	C. jejuni	0		Negative	0	0
10843179	23/03/06	11/07/15	1	169	1	3397	0	0	0		0	Salmonella	Negative	0	0
10843876	6/05/12	14/07/15	1	123	1	1164	1	0	0	C. jejuni	0		C. jejuni	1	0
10850449	3/09/14	5/08/15	1	53	1	336	0	0	0		0		Negative	0	0
10852174	26/04/12	12/08/15	1	125	1	1203	0	0	0		0	Shigella sonnei	Negative	0	0
10860835	14/09/14	7/09/15	1	57	1	358	0	0	0		0	Cl. difficile	Negative	0	0
10865636	10/07/13	22/09/15	1	109	1	804	1	0	0	C. jejuni	0		C. jejuni	1	0
10871847	5/05/15	9/10/15	1	19	0	157	0	0	0		0	Cl. difficile	Negative	0	0
10872379	5/05/15	9/10/15	1	18	0	157	0	0	0		0	Cl. difficile	Negative	0	0
10873919	7/10/13	15/10/15	1	102	1	738	0	0	0		0		Negative	0	0
10875253	22/11/14	19/10/15	1	52	0	331	0	0	0		0	Cl. difficile	Negative	0	0
10875847	11/12/13	21/10/15	1	97	1	679	0	0	0		0		Negative	0	0
10876106	8/09/14	21/10/15	1	64	1	408	1	0	0	C. jejuni	0		C. jejuni	1	0
10876938	30/11/14	24/10/15	1	50	0	328	0	0	0		0	Norovirus	Negative	0	0
10876958	24/08/15	24/10/15	1	6	1	61	0	0	0		0		Negative	0	0
10877075	17/09/15	25/10/15	1	3	1	38	0	0	0		0		Negative	0	0
10877782	6/10/15	27/10/15	1	1	1	21	0	0	0		0		Negative	0	0
10878239	4/05/15	27/10/15	1	26	1	176	0	0	0		0	Cl. difficile	Negative	0	0
10878247	27/10/10	27/10/15	1	143	1	1826	0	0	0		0	Cl. difficile	C. concisus GS2	0	1
10878911	18/05/15	29/10/15	1	21	0	164	0	0	0		0	Cl. difficile	Negative	0	0
10879449	6/02/14	30/10/15	1	92	1	631	0	0	0		0	Norovirus	Negative	0	0
10881394	15/03/12	5/11/15	1	130	1	1330	0	0	0		0	Adenovirus	C. concisus	0	1
10881824	7/01/15	6/11/15	1	46	1	303	0	1	0	C. concisus	0	Cl. difficile	Negative	0	0
10882525	17/11/13	9/11/15	0	102	1	722	0	0	0		0		Negative	0	0
10883036	9/12/10	10/11/15	1	143	1	1797	0	0	0		0		Negative	0	0
10883719	12/10/14	12/11/15	0	53	0	396	0	0	0		0		Negative	0	0
10884038	17/11/14	13/11/15	0	46	1	361	0	0	0		0	Cl. difficile	Negative	0	0
10884121	26/09/15	14/11/15	0	1	0	49	0	0	0		0		Negative	0	0
10884152	14/04/13	15/11/15	1	118	1	945	0	0	0		0	Giardia	Negative	0	0
10886012	28/08/14	18/11/15	1	70	0	447	0	0	0		0	Cl. difficile	Negative	0	0

Supplementary File 3

Sample Number	Date of Birth	Date of Collection	Case (1) or Control (0)	Matched Set	Gender (1=male)	Age (days)	C. jejuni	C. concisus	C. coli	Campylobactereaceae by Culture	Rotavirus	Other Taxa	Faecal MLPA Result	FaecalMLPA Jejuni	FaecalMLPA Concisus
10886439	11/08/14	19/11/15	0	76	0	465	0	0	0		0		C. concisus GS2	0	1
10886476	2/03/14	19/11/15	0	97	0	627	0	0	0		0		Negative	0	0
10886480	29/08/10	19/11/15	0	148	1	1908	0	0	0		0		Negative	0	0
10887180	12/06/14	21/11/15	1	80	1	527	1	0	0	C. jejuni	0		C. jejuni	1	0
10888154	9/06/13	25/11/15	1	115	1	899	0	0	0		0	Salmonella	C. concisus GS2 + C. ureolyticus	0	1
10888497	13/09/14	25/11/15	1	69	0	438	0	0	0		0		Negative	0	0
10889244	29/06/09	24/11/15	1	154	0	2339	0	0	0		0		Negative	0	0
10889338	18/12/10	25/11/15	1	144	1	1803	0	0	0		0	Dientamoeba	Negative	0	0
10889542	29/09/15	30/11/15	0	3	1	62	0	0	0		0		Negative	0	0
10889867	21/05/15	1/12/15	0	21	0	194	0	0	0		0		Negative	0	0
10890586	12/06/12	30/11/15	0	128	0	1266	0	0	0		0		Negative	0	0
10890924	31/05/15	2/12/15	0	24	0	185	0	0	0		0		Negative	0	0
10891414	21/05/14	3/12/15	0	80	0	561	0	0	0		NA		Negative	0	0
10892868	15/08/15	8/12/15	1	13	0	115	0	0	0		0		Negative	0	0
10892872	25/06/14	7/12/15	1	82	0	530	0	0	0		0		Negative	0	0
10892880	28/09/15	8/12/15	1	7	1	71	0	0	0		0	Adenovirus	Negative	0	0
10892896	22/11/14	7/12/15	1	58	0	380	0	0	0		0		Negative	0	0
10892911	21/10/14	8/12/15	1	66	0	413	0	0	0		0		C. concisus	0	1
10892916	6/03/13	8/12/15	0	122	1	1007	0	0	0		0		Negative	0	0
10892920	12/02/15	8/12/15	0	50	0	299	0	0	0		0	Cl. difficile	Negative	0	0
10893114	12/02/15	9/12/15	0	52	1	300	0	0	0		0		Negative	0	0
10895046	22/05/14	14/12/15	0	92	1	571	0	0	0		0	Cl. difficile	Negative	0	0
10895390	15/07/14	16/12/15	0	86	0	519	0	0	0		0	Cl. difficile	Negative	0	0
10896326	29/04/14	18/12/15	1	90	1	598	0	0	0		0	Cl. difficile	Negative	0	0
10896334	22/11/13	16/12/15	0	109	1	754	0	0	0		0	Dientamoeba	Negative	0	0
10896339	20/01/14	17/12/15	0	105	1	696	0	0	0		0	Blastocystis hominis	Negative	0	0
10896340	19/08/15	17/12/15	0	6	1	120	0	0	0		0		Negative	0	0
10896860	9/11/15	18/12/15	0	7	0	39	0	0	0		0		Negative	0	0
10896864	22/05/15	19/12/15	1	33	1	211	0	0	0		0	Cl. difficile	Negative	0	0
10896866	10/05/14	21/12/15	0	90	0	590	0	0	0		0		Negative	0	0
10899028	3/08/14	28/12/15	0	82	1	512	0	0	0		0		Negative	0	0
10899038	25/05/09	29/12/15	1	156	1	2409	1	0	0	C. jejuni	0		Negative	0	0
10899045	24/09/15	28/12/15	0	13	1	95	0	0	0		0		Negative	0	0
10899274	6/02/12	29/12/15	1	133	1	1422	0	0	0		0	Adenovirus	Negative	0	0
10899444	30/10/13	30/12/15	1	108	1	791	1	0	0	C. jejuni	0		Negative	0	0
10899991	8/07/15	31/12/15	0	25	0	176	0	0	0		0		Negative	0	0
10900315	24/04/12	4/01/16	0	130	1	1350	0	0	0		0		Negative	0	0
10900517	30/12/14	4/01/16	0	57	1	370	0	0	0		0		Negative	0	0
10901451	11/02/15	6/01/16	1	51	1	329	0	0	0		0	Cl. difficile	C. concisus GS2	0	1

Supplementary File 3

Sample Number	Date of Birth	Date of Collection	Case (1) or Control (0)	Matched Set	Gender (1=male)	Age (days)	C. jejuni	C. concisus	C. coli	Campylobactereaceae by Culture	Rotavirus	Other Taxa	Faecal MLPA Result	FaecalMLPA Jejuni	FaecalMLPA Concisus
10901498	13/09/15	6/01/16	1	14	1	115	0	0	0		0		Negative	0	0
10902380	6/07/09	8/01/16	0	156	1	2377	0	0	0		0		Campylobacter	0	0
10902707	19/01/15	9/01/16	1	56	1	355	0	0	0		1	Rorovirus	Negative	0	0
10903425	11/09/15	11/01/16	1	15	1	122	0	0	0		1	Rorovirus	Negative	0	0
10903484	20/08/14	11/01/16	1	77	0	509	0	0	0		0	Cl. difficile	Negative	0	0
10903624	26/12/13	11/01/16	1	103	0	746	0	0	0		1	Adenovirus, Rorovirus	Negative	0	0
10904112	8/03/09	13/01/16	0	159	0	2502	0	0	0		0		Negative	0	0
10904940	14/05/14	15/01/16	0	96	1	611	1	0	0	C. jejuni	0		Negative	0	0
10905169	14/04/14	17/01/16	1	94	0	643	0	0	0		0	Salmonella, Aeromonas hydrophila/caviae	Negative	0	0
10905238	19/01/14	18/01/16	1	100	1	729	0	0	0		0		Negative	0	0
10905433	26/03/11	16/01/16	1	141	0	1757	0	0	0		0	Adenovirus	Negative	0	0
10905552	14/04/12	18/01/16	0	133	0	1374	0	0	0		0		Negative	0	0
10905950	9/09/14	19/01/16	0	69	0	497	0	0	0		0		Not undertaken	NA	NA
10906550	18/12/13	17/01/16	1	106	1	760	0	0	0		0		Negative	0	0
10907093	8/02/12	21/01/16	1	134	0	1443	0	1	0	C. concisus	1	Rorovirus	C. concisus GS2	0	1
10907233	26/03/15	21/01/16	1	45	0	301	0	0	0		NA		Negative	0	0
10907794	18/08/14	23/01/16	1	79	0	523	0	0	0		1	Rorovirus	C. concisus GS2	0	1
10907798	5/10/10	23/01/16	1	147	0	1936	0	0	0		1	Dientamoeba, Rorovirus	Negative	0	0
10907867	31/05/12	20/01/16	1	129	0	1329	0	0	0		0		C. concisus GS2	0	1
10908156	27/08/14	25/01/16	0	77	1	516	0	0	0		0		Negative	0	0
10908612	2/11/15	26/01/16	1	10	0	85	0	0	0		0		Negative	0	0
10908631	6/04/11	26/01/16	1	141	1	1756	0	0	0		0		Negative	0	0
10908662	10/10/15	25/01/16	0	14	1	107	0	0	0		0		Negative	0	0
10908919	28/10/10	25/01/16	1	146	0	1915	0	0	0		0	Dientamoeba	Negative	0	0
10909014	7/10/14	27/01/16	0	70	1	477	0	0	0		0		Negative	0	0
10910216	18/04/11	29/01/16	1	140	0	1747	0	0	0		1	Rorovirus	Negative	0	0
10910282	29/06/14	29/01/16	1	88	0	579	0	0	0		0	Norovirus	Negative	0	0
10910285	14/04/13	31/01/16	1	121	1	1022	0	0	0		1	Giardia, Rorovirus	Negative	0	0
10910508	2/04/14	1/02/16	0	94	1	670	0	0	0		0	Norovirus	Negative	0	0
10910603	10/02/15	1/02/16	0	45	0	356	0	0	0		0		Negative	0	0
10911173	13/09/14	2/02/16	0	79	1	507	0	0	0		0	Cl. difficile	C. ureolyticus	0	0

Supplementary File 3

Sample Number	Date of Birth	Date of Collection	Case (1) or Control (0)	Matched Set	Gender (1=male)	Age (days)	C. jejuni	C. concisus	C. coli	Campylobactereaceae by Culture	Rotavirus	Other Taxa	Faecal MLPA Result	FaecalMLPA Jejuni	FaecalMLPA Concisus
10911195	21/08/15	2/02/16	0	15	1	165	0	0	0		0		Negative	0	0
10911395	4/06/09	1/02/16	1	157	0	2433	0	0	0		0	Salmonella	C. concisus GS2 + C. ureolyticus	0	1
10911696	16/05/15	3/02/16	0	33	0	263	0	0	0		0		Negative	0	0
10911698	13/03/14	3/02/16	0	103	0	692	1	0	0	C. jejuni	0	Cl. difficile	C. jejuni	1	0
10911699	13/03/14	3/02/16	0	100	0	692	0	0	0		0	Cl. difficile	Negative	0	0
10912010	13/03/14	4/02/16	0	94	1	693	0	0	0		0	Cl. difficile	C. concisus	0	1
10912014	24/02/14	4/02/16	0	106	0	710	0	0	0		0	Aeromonas caviae	Negative	0	0
10912052	12/01/11	4/02/16	0	143	1	1849	0	0	0		0		Negative	0	0
10912976	3/09/14	5/02/16	1	78	1	520	0	0	0		0		Negative	0	0
10912987	24/06/14	8/02/16	1	89	0	594	0	0	0		0	Yersinia enterocolitica	Negative	0	0
10913034	30/07/15	8/02/16	1	30	1	193	0	0	0		0	Cl. difficile	Negative	0	0
10913483	25/08/14	9/02/16	0	78	1	533	0	0	0		0	Cl. difficile	Negative	0	0
10913748	31/03/12	9/02/16	0	134	0	1410	0	0	0		0		Negative	0	0
10914117	21/10/13	10/02/16	0	108	0	842	0	0	0		0		Not undertaken	NA	NA
10914509	24/03/11	11/02/16	0	140	1	1785	0	0	0		0		Negative	0	0
10915008	30/07/13	11/02/16	1	116	1	926	0	0	0		0		Negative	0	0
10915229	10/06/12	12/02/16	1	131	1	1342	0	0	0		1	Rorovirus	Negative	0	0
10915770	27/02/15	15/02/16	0	51	0	353	0	0	0		0	Aeromonas caviae, Cl. Difficile, Dientamoeba	Negative	0	0
10915804	12/08/15	15/02/16	1	28	1	187	0	0	0		0		Negative	0	0
10916107	26/12/15	16/02/16	0	10	1	52	0	0	0		0		Negative	0	0
10916294	7/06/13	16/02/16	0	118	0	984	0	0	0		0		Negative	0	0
10916327	9/07/14	16/02/16	0	88	0	587	0	0	0		0		Negative	0	0
10916623	12/04/13	17/02/16	0	121	1	1041	0	0	0		0		Negative	0	0
10916981	8/11/12	18/02/16	1	124	0	1197	0	1	0	C. concisus	0		C. concisus GS2	0	1
10916987	7/08/15	18/02/16	0	26	1	195	0	0	0		0		C. concisus	0	1
10916993	28/01/15	17/02/16	1	55	0	385	0	0	0		0	Cl. difficile	Negative	0	0
10917005	3/05/09	18/02/16	0	157	0	2482	0	0	0		0		Negative	0	0
10917242	12/11/14	18/02/16	1	71	1	463	0	0	0		0		Negative	0	0
10917775	27/09/10	19/02/16	0	147	0	1971	0	0	0		0		Negative	0	0
10918669	6/08/15	23/02/16	0	28	1	201	0	0	0		0		Negative	0	0
10918671	2/10/15	22/02/16	1	17	1	143	0	0	0		0	Aeromonas hydrophila	Negative	0	0
10918793	14/05/14	23/02/16	1	95	0	650	0	0	0		0		Negative	0	0
10919824	18/11/11	21/02/16	1	135	1	1556	0	0	0		0	Salmonella	Negative	0	0
10920718	15/09/06	29/02/16	0	169	1	3454	0	0	0		0		Negative	0	0
10921300	3/12/10	1/03/16	0	146	1	1915	0	0	0		0		Negative	0	0
10921373	16/08/13	1/03/16	0	115	0	928	0	0	0		0	Giardia	Negative	0	0

Supplementary File 3

Sample Number	Date of Birth	Date of Collection	Case (1) or Control (0)	Matched Set	Gender (1=male)	Age (days)	C. jejuni	C. concisus	C. coli	Campylobactereaceae by Culture	Rotavirus	Other Taxa	Faecal MLPA Result	FaecalMLPA Jejuni	FaecalMLPA Concisus
10921992	28/09/15	2/03/16	0	17	1	156	0	0	0		0	Cl. difficile	C. concisus	0	1
10922083	7/01/16	2/03/16	1	5	1	55	1	0	0	C. jejuni	0		C. jejuni	1	0
10922840	9/07/12	4/03/16	0	129	0	1334	0	0	0		0		C. concisus	0	1
10923701	21/05/15	7/03/16	0	43	0	291	0	0	0		0	Cl. difficile	Negative	0	0
10923789	23/01/15	7/03/16	1	65	0	409	0	0	0		1	Rorovirus, Cl. difficile	C. concisus	0	1
10923792	31/01/05	7/03/16	0	179	1	4053	0	0	0		0		Negative	0	0
10924167	12/12/13	26/02/16	1	110	1	806	0	0	0		1	Rorovirus	C. concisus	0	1
10924790	27/01/14	8/03/16	1	107	0	771	0	0	0		1	Rorovirus	Negative	0	0
10925307	29/07/12	10/03/16	0	131	1	1320	0	0	0		0		Negative	0	0
10925990	30/06/08	12/03/16	1	163	1	2812	0	0	0		0	Blastocystis hominis, Norovirus	C. concisus	0	1
10927618	30/09/11	17/03/16	0	136	1	1630	0	0	0		0		Negative	0	0
10927737	1/12/14	16/03/16	1	73	1	471	0	0	0		0		Negative	0	0
10927956	15/04/14	17/03/16	1	98	0	702	0	0	0		1	Rorovirus	Negative	0	0
10928661	23/10/09	21/03/16	0	154	1	2341	0	1	0	C. concisus	0		Negative	0	0
10928732	26/06/05	21/03/16	1	177	0	3921	0	0	0		0	Salmonella	Negative	0	0
10928862	14/12/14	21/03/16	0	65	0	463	0	0	0		0		Negative	0	0
10930836	26/08/15	29/03/16	1	35	1	216	0	0	0		1	E. coli O111, Rorovirus	Negative	0	0
10930920	24/12/13	29/03/16	0	107	0	826	0	0	0		0		Negative	0	0
10932176	30/05/15	31/03/16	1	47	1	306	0	0	0		1	Rorovirus	Negative	0	0
10932730	16/11/12	4/04/16	0	124	1	1235	0	0	0		0		Negative	0	0
10934307	3/11/13	7/04/16	0	117	1	886	0	0	0		0		Negative	0	0
10934315	29/06/14	7/04/16	0	89	1	648	0	0	0		0		Negative	0	0
10934320	18/05/08	7/04/16	0	164	0	2881	0	0	0		0		Negative	0	0
10934665	12/03/15	8/04/16	0	55	1	393	0	0	0		0		Negative	0	0
10937065	6/08/15	14/04/16	1	40	0	252	0	0	0		0	Adenovirus	Negative	0	0
10937847	27/01/16	17/04/16	0	5	1	81	0	0	0		0		Negative	0	0
10938313	26/02/16	19/04/16	0	11	0	53	0	0	0		0		Negative	0	0
10938383	27/07/15	19/04/16	0	35	0	267	0	0	0		0	Cl. difficile	Negative	0	0
10938384	1/04/15	18/04/16	1	59	0	383	0	0	0		0		Negative	0	0
10939109	19/09/15	20/04/16	0	36	0	214	0	0	0		0		C. concisus GS2	0	1
10939423	21/02/15	18/04/16	0	58	1	422	0	0	0		0		C. concisus	0	1
10939944	2/09/11	21/04/16	0	139	1	1693	0	0	0		0		Negative	0	0
10940091	20/04/09	22/04/16	1	160	1	2559	1	0	0	C. jejuni	0		C. jejuni	1	0
10940118	8/11/11	21/04/16	1	138	0	1626	0	0	0		1	Rorovirus	Negative	0	0
10940182	21/10/15	21/04/16	1	27	1	183	0	0	0		0		Negative	0	0
10940507	18/04/14	25/04/16	0	98	0	738	1	0	0	C. jejuni	0		C. concisus GS2	0	1
10940716	29/06/15	25/04/16	0	40	0	301	0	0	0		0		Negative	0	0

Supplementary File 3

Sample Number	Date of Birth	Date of Collection	Case (1) or Control (0)	Matched Set	Gender (1=male)	Age (days)	C. jejuni	C. concisus	C. coli	Campylobactereaceae by Culture	Rotavirus	Other Taxa	Faecal MLPA Result	FaecalMLPA Jejuni	FaecalMLPA Concisus
10941593	24/12/14	26/04/16	1	74	0	489	0	0	0		0		C. concisus GS2	0	1
10941826	18/12/05	27/04/16	1	175	0	3783	0	0	0		0		Negative	0	0
10942366	29/09/10	29/04/16	0	150	0	2039	0	0	0		0		Negative	0	0
10942644	11/08/05	29/04/16	0	177	1	3914	0	0	0		0	Cl. difficile	C. concisus	0	1
10942736	16/12/14	29/04/16	1	75	1	500	1	0	0	C. jejuni	0		Negative	0	0
10943202	29/04/11	2/05/16	0	144	1	1830	0	0	0		0	Norovirus	C. concisus	0	1
10945089	7/03/15	6/05/16	1	68	0	426	0	0	0		0	Cl. difficile	Negative	0	0
10945093	1/10/13	6/05/16	1	119	0	948	0	0	0		0	Norovirus	Negative	0	0
10945436	14/04/15	9/05/16	1	59	1	391	0	0	0		0	Cl. difficile	C. concisus	0	1
10945895	19/08/15	10/05/16	0	47	1	265	0	0	0		0	Cl. difficile	Negative	0	0
10947293	17/12/15	13/05/16	0	27	1	148	0	0	0		0	Cl. difficile	Negative	0	0
10947489	6/10/15	13/05/16	1	37	1	220	0	0	0		0		C. concisus GS2	0	1
10947883	29/12/07	17/05/16	0	167	1	3062	0	0	0		0	Cl. difficile	Negative	0	0
10948495	5/11/13	18/05/16	0	119	0	925	0	0	0		0	Blastocystis hominis, Norovirus	Negative	0	0
10948818	28/11/11	18/05/16	0	138	1	1633	0	0	0		0		Negative	0	0
10949149	2/05/10	19/05/16	0	152	1	2209	0	0	0		0		Negative	0	0
10949912	14/07/15	20/05/16	0	56	1	311	0	0	0		0	Cl. difficile	Negative	0	0
10949919	4/12/15	20/05/16	1	22	1	168	0	0	0		0	E. coli O111	Negative	0	0
10950378	7/04/15	23/05/16	0	59	1	412	0	0	0		0	Cl. difficile	Negative	0	0
10951449	5/02/15	25/05/16	0	68	0	475	0	0	0		0	Cl. difficile	Negative	0	0
10951760	21/01/13	26/05/16	1	126	0	1221	0	1	0	C. concisus	0	Salmonella	C. concisus GS1 & GS2	0	1
10952105	18/11/15	27/05/16	1	29	1	191	0	0	0		0		Negative	0	0
10952514	22/02/15	25/05/16	0	66	0	458	0	1	0	C. concisus	0	Cl. difficile	Negative	0	0
10953134	18/03/16	30/05/16	0	2	0	73	0	0	0		0		Negative	0	0
10953803	9/08/11	1/06/16	0	142	0	1758	0	0	0		0		Negative	0	0
10953993	22/12/12	1/06/16	1	127	1	1257	0	0	0		0	Blastocystis hominis	Negative	0	0
10955153	8/02/14	3/06/16	0	110	0	846	0	0	0		0		C. concisus GS2	0	1
10955729	17/01/14	6/06/16	1	113	1	871	1	0	0	C. jejuni	0	Aeromonas caviae	Negative	0	0
10957026	18/10/05	9/06/16	0	177	0	3887	0	0	0		0		Negative	0	0
10957317	1/04/09	9/06/16	0	161	0	2626	0	0	0		0	Dientamoeba	Negative	0	0
10957350	19/11/14	9/06/16	1	84	0	568	0	0	0		1	Rorovirus	Negative	0	0
10957573	24/02/15	10/06/16	0	67	1	472	0	0	0		0	Cl. difficile	Negative	0	0
10957577	15/05/14	10/06/16	0	99	0	757	0	0	0		0		Negative	0	0

Supplementary File 3

Sample Number	Date of Birth	Date of Collection	Case (1) or Control (0)	Matched Set	Gender (1=male)	Age (days)	C. jejuni	C. concisus	C. coli	Campylobactereaceae by Culture	Rotavirus	Other Taxa	Faecal MLPA Result	FaecalMLPA Jejuni	FaecalMLPA Concisus
10958138	3/10/15	13/06/16	0	37	1	254	0	0	0		0	Cl. difficile	Negative	0	0
10958575	13/02/15	14/06/16	0	71	0	487	NA	NA	NA	Not undertaken	0	E. coli O126	Negative	0	0
10958583	24/01/14	14/06/16	0	117	1	872	NA	NA	NA	Not undertaken	0		Negative	0	0
10959457	21/08/15	16/06/16	0	48	1	300	0	0	0		0	Cl. difficile	Not undertaken	NA	NA
10960053	23/02/15	17/06/16	0	73	1	480	0	0	0		0	Aeromonas caviae	Negative	0	0
10962903	4/03/14	24/06/16	0	113	1	843	0	0	0		0		Negative	0	0
10967666	4/04/14	5/07/16	1	111	1	823	1	0	0	C. jejuni	NA	Cl. difficile	C. jejuni	1	0
10974083	22/12/15	20/07/16	1	34	1	211	0	0	0		0		Negative	0	0
10974159	10/08/15	21/07/16	1	54	1	346	0	0	0		0	Cl. difficile	Negative	0	0
10974735	27/04/16	22/07/16	1	9	0	86	0	0	0		0	Cl. difficile	Negative	0	0
10974997	31/01/08	25/07/16	0	168	0	3098	0	0	0		0		Negative	0	0
10975269	10/03/16	25/07/16	0	9	1	137	0	0	0		0		Negative	0	0
10975280	23/11/14	25/07/16	0	84	0	610	0	1	0	C. concisus	0		Negative	0	0
10976591	22/10/15	28/07/16	1	42	0	280	0	0	0		0		C. concisus GS2	0	1
10980405	28/06/14	8/08/16	0	101	1	772	0	0	0		0		Negative	0	0
10981895	16/05/16	11/08/16	0	8	1	87	0	0	0		0		Negative	0	0
10982518	25/02/16	12/08/16	0	12	1	169	0	0	0		0		C. concisus GS2	0	1
10982519	30/07/14	12/08/16	0	104	1	744	0	0	0		0		Negative	0	0
10983181	5/12/15	15/08/16	1	41	0	254	1	0	0	C. jejuni	0	Dientamoeba	C. jejuni	1	0
10984299	1/10/14	17/08/16	0	95	1	686	0	0	0		0	Cl. difficile	Negative	0	0
10984553	9/04/14	18/08/16	0	114	1	862	0	1	0	C. concisus	0		Negative	0	0
10984714	27/10/14	18/08/16	0	91	1	661	0	0	0		0		Negative	0	0
10985240	25/12/15	16/08/16	0	29	1	235	0	0	0		0	Cl. difficile	Negative	0	0
10985260	16/12/14	19/08/16	0	85	1	612	0	0	0		0		Negative	0	0
10985295	7/05/14	19/08/16	0	113	1	835	0	0	0		0		Negative	0	0
10985297	7/05/14	19/08/16	0	111	1	835	0	0	0		0		Negative	0	0
10985743	3/12/06	22/08/16	0	172	1	3550	0	0	0		0		Negative	0	0
10985746	10/12/15	22/08/16	0	32	0	256	0	0	0		0		C. concisus	0	1
10986453	27/12/15	23/08/16	0	30	1	240	0	0	0		0		Negative	0	0
10987715	19/12/12	25/08/16	0	132	1	1345	0	0	0		0	Giardia	Negative	0	0
10988481	10/02/16	25/08/16	1	20	0	197	0	0	0		0		Negative	0	0
10989941	12/05/15	30/08/16	0	74	0	476	0	0	0		0		Negative	0	0
10996219	20/12/15	13/09/16	0	34	1	268	0	0	0		0		Not undertaken	NA	NA
10997652	20/03/15	15/09/16	0	75	0	545	0	0	0		0		Negative	0	0
10997667	27/07/15	15/09/16	0	60	1	416	0	0	0		0		Negative	0	0
10997676	1/02/15	15/09/16	0	87	1	592	0	0	0		0		Negative	0	0
10999803	27/05/07	21/09/16	0	170	1	3405	0	0	0		0	E. coli O111	Negative	0	0
11000531	13/12/15	16/09/16	0	39	0	278	0	0	0		0	Cl. difficile	Negative	0	0
11000549	13/12/15	16/09/16	0	41	0	278	0	0	0		0	Cl. difficile	Negative	0	0
11000618	31/08/05	21/09/16	0	178	0	4039	0	0	0		0		Negative	0	0
11000841	20/02/16	22/09/16	0	20	1	215	0	0	0		0		Negative	0	0

Supplementary File 3

Sample Number	Date of Birth	Date of Collection	Case (1) or Control (0)	Matched Set	Gender (1=male)	Age (days)	C. jejuni	C. concisus	C. coli	Campylobactereaceae by Culture	Rotavirus	Other Taxa	Faecal MLPA Result	FaecalMLPA Jejuni	FaecalMLPA Concisus
11000898	1/12/13	22/09/16	0	120	1	1026	0	0	0	A. butzleri	0		Negative	0	0
11001332	19/12/05	23/09/16	0	176	1	3931	0	0	0		0		Negative	0	0
11003530	4/05/16	28/09/16	0	16	0	147	0	0	0		0		Negative	0	0
11003706	31/08/07	28/09/16	0	171	0	3316	0	0	0		0		Negative	0	0
11003917	9/07/15	29/09/16	0	63	1	448	0	0	0		0		Negative	0	0
11005462	24/09/15	3/10/16	0	54	0	375	0	0	0		0	Cl. difficile	Negative	0	0
11005535	2/04/09	3/10/16	0	162	0	2741	0	0	0		0		Negative	0	0
11006998	18/09/12	5/10/16	0	134	0	1478	0	0	0		0		C. concisus GS1 & GS2	0	1
11008004	5/10/15	31/08/16	0	42	0	331	0	0	0		0	Cl. difficile	Negative	0	0
17428547	6/11/14	25/05/16	1	87	0	566	0	0	0		0		C. concisus	0	1
80980182	19/09/14	7/05/15	1	38	1	230	0	0	0		0		Negative	0	0
80981067	30/07/08	11/05/15	1	158	0	2476	0	0	0		0		Negative	0	0
80981998	30/10/14	14/05/15	1	31	1	196	1	0	0	C. jejuni	0		Negative	0	0
80982078	21/07/14	15/05/15	1	44	1	298	1	0	0	C. jejuni	0	Adenovirus	Not undertaken	NA	NA
80982109	22/10/14	15/05/15	0	23	1	205	0	0	0		0	E. coli O114, Cl. Difficile	Negative	0	0
80982208	28/12/13	15/05/15	0	81	0	503	0	0	0		0		Not undertaken	NA	NA
80982211	22/07/10	15/05/15	0	145	1	1758	0	0	0		0		Not undertaken	NA	NA
80983034	3/01/14	19/05/15	1	76	1	501	0	0	0		0	Adenovirus	Not undertaken	NA	NA
80985548	31/01/09	29/05/15	0	155	0	2309	0	0	0		0		Not undertaken	NA	NA
80985613	19/07/08	29/05/15	0	158	0	2505	0	0	0		0		Not undertaken	NA	NA
80985689	21/10/14	29/05/15	0	38	1	220	0	0	0		0	Cl. difficile	Negative	0	0
80985714	28/04/15	29/05/15	0	4	0	31	0	0	0		0		Not undertaken	NA	NA
80985815	16/12/14	29/05/15	0	31	1	164	0	0	0		0		Not undertaken	NA	NA
80986325	7/06/14	2/06/15	0	61	0	360	0	0	0		0		Not undertaken	NA	NA
80987173	22/12/10	3/06/15	1	136	1	1624	0	0	0		1	Rorovirus	Negative	0	0
80991380	25/12/14	19/06/15	1	25	0	176	0	0	0		NA		Not undertaken	NA	NA
80995439	25/01/13	6/07/15	1	114	0	892	0	0	0		0		Negative	0	0
80998795	9/02/11	17/07/15	0	136	1	1619	0	0	0		0		Negative	0	0
80998837	8/10/13	17/07/15	0	93	0	647	0	0	0		0		Negative	0	0
80999518	20/02/13	23/07/15	0	114	0	883	0	0	0		0		Negative	0	0
81001292	21/03/05	30/07/15	1	174	0	3783	0	0	0		0		Negative	0	0
81001909	24/06/14	3/08/15	0	62	0	405	0	0	0		0		Negative	0	0
81001910	22/07/12	3/08/15	0	123	0	1107	0	0	0		0		Not undertaken	NA	NA
81002107	12/07/12	3/08/15	0	NA	1	1117	NA	NA	NA	Not undertaken	0	Not undertaken	Not undertaken	NA	NA
81004068	29/04/14	10/08/15	1	72	0	468	0	0	0		0	Cl. difficile	Negative	0	0
81005629	4/07/14	17/08/15	0	NA	0	409	NA	NA	NA	Not undertaken	0	Not undertaken	Not undertaken	NA	NA
81005630	11/09/08	14/08/15	1	159	1	2528	0	0	0		0		Not undertaken	NA	NA
81006274	14/04/13	19/08/15	1	112	0	857	0	0	0		0		Negative	0	0
81009742	3/05/05	31/08/15	0	174	0	3772	0	0	0		0		Negative	0	0
81012499	21/05/13	11/09/15	0	112	1	843	0	0	0		0		Negative	0	0
81012500	2/03/14	11/09/15	0	83	0	558	0	0	0		0		Negative	0	0

Supplementary File 3

Sample Number	Date of Birth	Date of Collection	Case (1) or Control (0)	Matched Set	Gender (1=male)	Age (days)	C. jejuni	C. concisus	C. coli	Campylobactereaceae by Culture	Rotavirus	Other Taxa	Faecal MLPA Result	FaecalMLPA Jejuni	FaecalMLPA Concisus
81026075	10/11/08	2/11/15	0	159	0	2548	0	0	0		0	Yersinia enterocolitica	Negative	0	0
81026142	5/05/15	2/11/15	0	19	1	181	0	0	0		0	Cl. difficile	Negative	0	0
81026219	18/06/14	2/11/15	0	72	1	502	0	0	0		0		Negative	0	0
81026313	14/06/15	2/11/15	0	18	0	141	0	0	0		0		Not undertaken	NA	NA
81026531	6/04/06	3/11/15	0	173	0	3498	0	0	0		0		Negative	0	0
81026553	17/07/13	3/11/15	0	109	1	839	0	0	0		0		Negative	0	0
81026790	4/09/14	4/11/15	0	64	0	426	0	0	0		0		Negative	0	0
81027503	2/11/14	6/11/15	0	49	1	369	0	0	0		0	Aeromonas caviae, Cl. Difficile	Negative	0	0
81027744	12/10/13	6/11/15	1	105	0	755	0	0	0		0		Negative	0	0
81028359	23/04/14	8/11/15	1	86	0	564	0	0	0		0		Negative	0	0
81030211	29/01/14	13/11/15	1	96	0	653	1	0	0	C. jejuni	0	Cl. difficile	C. jejuni	1	0
81033505	5/12/14	27/11/15	0	44	1	357	0	0	0		0	Cl. difficile	Negative	0	0
81050641	26/10/15	26/01/16	1	11	1	92	0	0	0		0		Negative	0	0
81055035	24/02/15	11/02/16	1	55	0	352	0	0	0		0	E. coli O26	Negative	0	0
81055043	6/07/15	11/02/16	1	36	1	220	0	1	0	C. concisus	0	Norovirus, Cl. difficile	Not undertaken	NA	NA
81055176	7/06/14	8/02/16	1	91	0	611	1	0	0	C. jejuni	0		Not undertaken	NA	NA
81055223	27/04/15	11/02/16	1	43	1	290	0	0	0		0		Negative	0	0
81057685	4/08/06	19/02/16	1	169	0	3486	0	0	0		0		Negative	0	0
81058343	30/11/09	20/02/16	1	153	1	2273	1	0	0	C. jejuni	0		C. jejuni	1	0
81058472	20/01/05	20/02/16	1	179	1	4048	1	0	0	C. jejuni	0	Cl. difficile	C. jejuni	1	0
81059654	16/08/14	25/02/16	1	85	1	558	1	0	0	C. jejuni	0	Aeromonas caviae	Negative	0	0
81060271	29/04/14	25/02/16	1	95	0	667	1	0	0	C. jejuni	0	Aeromonas caviae	C. jejuni	1	0
81061452	17/08/13	3/03/16	0	116	1	929	0	0	0		0		Negative	0	0
81061562	13/12/13	3/03/16	0	110	1	811	0	0	1	C. coli	0		C. coli	0	0
81061807	25/03/11	7/03/16	0	141	1	1809	0	0	0		0		Negative	0	0
81065359	29/08/11	17/03/16	1	139	1	1662	0	0	0		0		Negative	0	0
81065939	16/11/07	18/03/16	1	167	1	3045	0	0	0		0	Blastocystis hominis, Dientamoeba	Negative	0	0
81067085	18/07/07	22/02/16	1	168	1	3141	0	0	0		0	Blastocystis hominis, Dientamoeba	Negative	0	0
81067880	8/09/13	25/03/16	1	117	1	929	0	0	0		1	Rorovirus	Negative	0	0
81074287	17/03/16	19/04/16	1	2	0	33	0	0	0		0	Cl. difficile	Negative	0	0
81075799	29/03/10	26/04/16	1	152	0	2220	0	0	0		0		Not undertaken	NA	NA
81076392	23/10/14	25/04/16	1	84	0	550	0	0	0		0	Cl. difficile	Negative	0	0
81076609	7/04/14	29/04/16	1	99	0	753	0	0	0		0	Aeromonas hydrophila/caviae	Negative	0	0
81078259	12/04/10	6/05/16	1	152	0	2216	0	0	0		0		Negative	0	0

Supplementary File 3

Sample Number	Date of Birth	Date of Collection	Case (1) or Control (0)	Matched Set	Gender (1=male)	Age (days)	C. jejuni	C. concisus	C. coli	Campylobactereaceae by Culture	Rotavirus	Other Taxa	Faecal MLPA Result	FaecalMLPA Jejuni	FaecalMLPA Concisus
81078872	10/06/11	7/05/16	1	142	0	1793	1	0	0	C. jejuni	0		C. concisus + C. jejuni	1	1
81079425	23/04/15	10/05/16	1	60	0	383	0	0	0		0	Cl. difficile	Negative	0	0
81081451	28/03/15	15/05/16	1	67	1	414	0	0	0		0	Cl. difficile	C. concisus GS2	0	1
81081631	1/03/16	17/05/16	1	9	1	77	0	0	0		0	Norovirus, Cl. difficile	Negative	0	0
81082674	18/05/14	20/05/16	1	101	0	733	0	0	0		0		Negative	0	0
81085323	14/06/14	30/05/16	1	99	1	716	0	0	0		0	Giardia	Negative	0	0
81087854	6/05/15	7/06/16	1	63	0	398	0	0	0		0	Adenovirus	Negative	0	0
81088350	26/09/14	8/06/16	1	85	1	621	0	0	0		0		C. jejuni	1	0
81088435	13/11/15	9/06/16	0	22	0	209	0	0	0		0		Negative	0	0
81088706	31/07/15	9/06/16	1	48	1	314	0	0	0		0	Norovirus	Negative	0	0
81090488	17/03/08	15/06/16	1	166	0	3012	0	0	0		0		Negative	0	0
81091372	7/04/16	18/06/16	1	8	1	72	0	0	0		0		Negative	0	0
81094123	1/12/15	28/06/16	1	32	1	210	1	0	0	C. jejuni	0	Aeromonas sp.	C. jejuni	1	0
81094482	21/10/13	30/06/16	1	120	1	983	0	0	0		0		Negative	0	0
81098106	6/09/14	13/07/16	1	91	1	676	0	0	0		0		Negative	0	0
81100344	9/11/12	22/07/16	1	132	1	1351	1	0	0	C. jejuni	0		C. jejuni	1	0
81101045	4/07/14	26/07/16	1	104	0	753	0	1	0	C. concisus	0		Negative	0	0
81102268	7/04/16	30/07/16	1	12	1	114	0	0	0		0	Cl. difficile	Negative	0	0
81102337	6/12/15	1/08/16	1	39	0	239	0	0	0		0	Cl. difficile	Negative	0	0
81102587	14/02/16	2/08/16	1	12	1	170	0	0	0		0	Cl. difficile	Negative	0	0
81102588	30/03/16	2/08/16	1	16	1	125	0	0	0		0		Negative	0	0
81103417	22/12/06	5/08/16	1	172	0	3514	0	0	0		0		Negative	0	0
81104380	4/03/16	10/08/16	1	20	0	159	0	0	0		0		Negative	0	0
81105862	30/10/15	17/08/16	1	NA	1	292	NA	NA	NA	Not undertaken	0	Not undertaken	Not undertaken	NA	NA
81106009	6/03/12	17/08/16	1	137	1	1625	0	0	0		0		Negative	0	0
81107814	1/12/05	24/08/16	1	176	0	3919	0	0	0		0		Negative	0	0
81110257	30/06/05	1/09/16	1	165	1	4081	0	0	0		0	Aerococcus sp.	Negative	0	0
81112309	15/11/10	10/09/16	1	151	0	2126	0	0	0		0		Negative	0	0
81113260	7/10/05	13/09/16	1	178	1	3994	0	0	0		0	Cryptosporidium	C. concisus GS2	0	1
81113568	22/05/07	15/09/16	1	170	1	3404	0	0	0		0		Negative	0	0
81114620	7/09/15	20/09/16	1	54	1	379	0	0	0		0		C. concisus GS2	0	1
81115274	16/03/09	21/09/16	1	162	1	2746	0	0	0		0		Negative	0	0
81115605	24/01/11	23/09/16	1	149	1	2069	0	0	0		0	Adenovirus	Negative	0	0

Supplementary File 3

Sample Number	FaecalMLPA GS1	FaecalMLPA GS2	FaecalTaxon-specificPCRResults	CompositeGold StandardResult	MLPA Score	CGS concisus	CGS concisusGS2	CGS jejuni	Isolate Name	Isolate MLPA Result	IsolateAdditionalResults	Discordant
10817726	0	0	Positive for Campylobacter and C. jejuni	C. jejuni	FN	0	0	1	Cam15/64	C. jejuni	Positive for C. jejuni	Yes
10818452	0	0		Negative	TN	0	0	0				No
10819077	0	0		Negative	TN	0	0	0				No
10820448	0	0		Negative	TN	0	0	0				No
10825751	NA	NA		NA	NA	NA	NA	NA				
10826643	NA	NA		NA	NA	NA	NA	NA				
10837941	NA	NA		NA	NA	NA	NA	NA				
10838427	NA	NA		NA	NA	NA	NA	NA				
10841829	0	0	Positive for C. jejuni	C. jejuni	TP	0	0	1	Cam15/87	C. jejuni	Positive for C. jejuni	No
10843131	0	0	Negative for Campylobacter, C. coli & C. jejuni	C. jejuni	FN	0	0	1	Cam15/88	C. jejuni	Positive for C. jejuni	Yes
10843179	0	0		Negative	TN	0	0	0				No
10843876	0	0	Positive for C. jejuni	C. jejuni	TP	0	0	1	Cam15/94	C. jejuni	Positive for C. jejuni	No
10850449	0	0		Negative	TN	0	0	0				No
10852174	0	0		Negative	TN	0	0	0				No
10860835	0	0		Negative	TN	0	0	0				No
10865636	0	0	Positive for C. jejuni	C. jejuni	TP	0	0	1	Cam15/152	C. jejuni	Positive for C. jejuni	No
10871847	0	0		Negative	TN	0	0	0				No
10872379	0	0		Negative	TN	0	0	0				No
10873919	0	0		Negative	TN	0	0	0				No
10875253	0	0		Negative	TN	0	0	0				No
10875847	0	0		Negative	TN	0	0	0				No
10876106	0	0	Positive for C. jejuni	C. jejuni	TP	0	0	1	Cam15/173	C. jejuni	Positive for C. jejuni	No
10876938	0	0		Negative	TN	0	0	0				No
10876958	0	0		Negative	TN	0	0	0				No
10877075	0	0		Negative	TN	0	0	0				No
10877782	0	0		Negative	TN	0	0	0				No
10878239	0	0		Negative	TN	0	0	0				No
10878247	0	1	Positive for Campylobacter, C. concisus and GS2	C. concisus GS2	TP	1	1	0				Yes
10878911	0	0	Negative for Campylobacter, C. cuniculorum, C. helveticus and C. upsaliensis	Negative	TN	0	0	0				No
10879449	0	0		Negative	TN	0	0	0				No
10881394	0	0	Positive for Campylobacter and GS2; negative for C. concisus	C. concisus	TP	1	0	0				Yes
10881824	0	0	Negative for Campylobacter, C. concisus and GS1 & GS2	C. concisus	FN	1	0	0	Cam15/223	C. concisus GS2	Positive for GS2; negative for C. concisus	Yes
10882525	0	0		Negative	TN	0	0	0				No
10883036	0	0		Negative	TN	0	0	0				No
10883719	0	0		Negative	TN	0	0	0				No
10884038	0	0		Negative	TN	0	0	0				No
10884121	0	0		Negative	TN	0	0	0				No
10884152	0	0		Negative	TN	0	0	0				No
10886012	0	0		Negative	TN	0	0	0				No

Supplementary File 3

Sample Number	FaecalMLPA GS1	FaecalMLPA GS2	FaecalTaxon-specificPCRResults	CompositeGold StandardResult	MLPA Score	CGS concisus	CGS concisusGS2	CGS jejuni	Isolate Name	Isolate MLPA Result	IsolateAdditionalResults	Discordant
10886439	0	1	Positive for Campylobacter, C. concisus and GS2	C. concisus GS2	TP	1	1	0				Yes
10886476	0	0		Negative	TN	0	0	0				No
10886480	0	0		Negative	TN	0	0	0				No
10887180	0	0	Positive for C. jejuni	C. jejuni	TP	0	0	1	Cam15/230	C. jejuni	Positive for C. jejuni	No
10888154	0	1	Positive for Campylobacter, C. concisus and GS2; negative for C. ureolyticus	C. concisus GS2	TP	1	1	0				Yes
10888497	0	0		Negative	TN	0	0	0				No
10889244	0	0		Negative	TN	0	0	0				No
10889338	0	0		Negative	TN	0	0	0				No
10889542	0	0		Negative	TN	0	0	0				No
10889867	0	0	Negative for Campylobacter, C. cuniculorum, C. helveticus and C. upsaliensis	Negative	TN	0	0	0				No
10890586	0	0		Negative	TN	0	0	0				No
10890924	0	0		Negative	TN	0	0	0				No
10891414	0	0		Negative	TN	0	0	0				No
10892868	0	0		Negative	TN	0	0	0				No
10892872	0	0		Negative	TN	0	0	0				No
10892880	0	0		Negative	TN	0	0	0				No
10892896	0	0		Negative	TN	0	0	0				No
10892911	0	0	Positive for GS2; negative for Campylobacter and C. concisus	C. concisus	TP	1	0	0				Yes
10892916	0	0		Negative	TN	0	0	0				No
10892920	0	0		Negative	TN	0	0	0				No
10893114	0	0		Negative	TN	0	0	0				No
10895046	0	0		Negative	TN	0	0	0				No
10895390	0	0		Negative	TN	0	0	0				No
10896326	0	0		Negative	TN	0	0	0				No
10896334	0	0		Negative	TN	0	0	0				No
10896339	0	0		Negative	TN	0	0	0				No
10896340	0	0		Negative	TN	0	0	0				No
10896860	0	0		Negative	TN	0	0	0				No
10896864	0	0		Negative	TN	0	0	0				No
10896866	0	0		Negative	TN	0	0	0				No
10899028	0	0		Negative	TN	0	0	0				No
10899038	0	0	Positive for C. jejuni; negative for Campylobacter	C. jejuni	FN	0	0	1	Cam15/259	C. jejuni	Positive for C. jejuni	Yes
10899045	0	0		Negative	TN	0	0	0				No
10899274	0	0		Negative	TN	0	0	0				No
10899444	0	0	Positive for Campylobacter and C. jejuni	C. jejuni	FN	0	0	1	Cam15/266	C. jejuni	Positive for C. jejuni	Yes
10899991	0	0		Negative	TN	0	0	0				No
10900315	0	0		Negative	TN	0	0	0				No
10900517	0	0		Negative	TN	0	0	0				No
10901451	0	1	Positive for Campylobacter, C. concisus and GS1	C. concisus GS2	TP	1	1	0				Yes

Supplementary File 3

Sample Number	FaecalMLPA GS1	FaecalMLPA GS2	FaecalTaxon-specificPCRResults	CompositeGold StandardResult	MLPA Score	CGS concisus	CGS concisusGS2	CGS jejuni	Isolate Name	Isolate MLPA Result	IsolateAdditionalResults	Discordant
10901498	0	0		Negative	TN	0	0	0				No
10902380	0	0	Positive for Campylobacter; negative for A. butzleri, C. coli & C. jejuni, C. concisus, GS1 & GS2, C. cuniculorum, C. helveticus, C. lari, C. upsaliensis and C. ureolyticus	Campylobacter	TP	0	0	0				Yes
10902707	0	0		Negative	TN	0	0	0				No
10903425	0	0		Negative	TN	0	0	0				No
10903484	0	0		Negative	TN	0	0	0				No
10903624	0	0		Negative	TN	0	0	0				No
10904112	0	0		Negative	TN	0	0	0				No
10904940	0	0	Positive for C. jejuni; negative for Campylobacter	C. jejuni	FN	0	0	1	Cam16/14	C. jejuni	Positive for C. jejuni	Yes
10905169	0	0		Negative	TN	0	0	0				No
10905238	0	0		Negative	TN	0	0	0				No
10905433	0	0		Negative	TN	0	0	0				No
10905552	0	0		Negative	TN	0	0	0				No
10905950	NA	NA		NA	NA	NA	NA	NA				
10906550	0	0		Negative	TN	0	0	0				No
10907093	0	1	Positive for C. concisus and GS2	C. concisus GS2	TP	1	0	0	Cam16/24	C. concisus GS1	Positive for C. concisus and GS1	Yes
10907233	0	0		Negative	TN	0	0	0				No
10907794	0	1	Positive for Campylobacter, C. concisus and GS2	C. concisus GS2	TP	1	1	0				Yes
10907798	0	0		Negative	TN	0	0	0				No
10907867	0	1	Positive for GS2; negative for Campylobacter and C. concisus	C. concisus GS2	TP	1	1	0				Yes
10908156	0	0		Negative	TN	0	0	0				No
10908612	0	0		Negative	TN	0	0	0				No
10908631	0	0		Negative	TN	0	0	0				No
10908662	0	0		Negative	TN	0	0	0				No
10908919	0	0		Negative	TN	0	0	0				No
10909014	0	0		Negative	TN	0	0	0				No
10910216	0	0		Negative	TN	0	0	0				No
10910282	0	0		Negative	TN	0	0	0				No
10910285	0	0		Negative	TN	0	0	0				No
10910508	0	0		Negative	TN	0	0	0				No
10910603	0	0		Negative	TN	0	0	0				No
10911173	0	0	Positive for Campylobacter and C. ureolyticus	C. ureolyticus	TP	0	0	0				Yes

Supplementary File 3

Sample Number	FaecalMLPA GS1	FaecalMLPA GS2	FaecalTaxon-specificPCRResults	CompositeGold StandardResult	MLPA Score	CGS concisus	CGS concisusGS2	CGS jejuni	Isolate Name	Isolate MLPA Result	IsolateAdditionalResults	Discordant
10911195	0	0		Negative	TN	0	0	0				No
10911395	0	1	Positive for Campylobacter, C. concisus and GS2; negative for C. ureolyticus	C. concisus GS2	TP	1	1	0				Yes
10911696	0	0		Negative	TN	0	0	0				No
10911698	0	0	Positive for Campylobacter; negative for C. coli & C. jejuni	C. jejuni	TP	0	0	1	Cam16/31	C. jejuni	Positive for C. jejuni	No
10911699	0	0		Negative	TN	0	0	0				No
10912010	0	0	Positive for Campylobacter, C. concisus and GS2	C. concisus	TP	1	0	0				Yes
10912014	0	0		Negative	TN	0	0	0				No
10912052	0	0		Negative	TN	0	0	0				No
10912976	0	0		Negative	TN	0	0	0				No
10912987	0	0		Negative	TN	0	0	0				No
10913034	0	0		Negative	TN	0	0	0				No
10913483	0	0		Negative	TN	0	0	0				No
10913748	0	0		Negative	TN	0	0	0				No
10914117	NA	NA		NA	NA	NA	NA	NA				
10914509	0	0		Negative	TN	0	0	0				No
10915008	0	0		Negative	TN	0	0	0				No
10915229	0	0		Negative	TN	0	0	0				No
10915770	0	0		Negative	TN	0	0	0				No
10915804	0	0		Negative	TN	0	0	0				No
10916107	0	0		Negative	TN	0	0	0				No
10916294	0	0		Negative	TN	0	0	0				No
10916327	0	0		Negative	TN	0	0	0				No
10916623	0	0		Negative	TN	0	0	0				No
10916981	0	1	Positive for C. concisus and GS1+GS2	C. concisus	TP	1	0	0	Cam16/52	Not undertaken		No
10916987	0	0	Positive for Campylobacter and GS2; negative for C. concisus	C. concisus	TP	1	0	0				Yes
10916993	0	0		Negative	TN	0	0	0				No
10917005	0	0		Negative	TN	0	0	0				No
10917242	0	0		Negative	TN	0	0	0				No
10917775	0	0		Negative	TN	0	0	0				No
10918669	0	0		Negative	TN	0	0	0				No
10918671	0	0		Negative	TN	0	0	0				No
10918793	0	0		Negative	TN	0	0	0				No
10919824	0	0		Negative	TN	0	0	0				No
10920718	0	0		Negative	TN	0	0	0				No
10921300	0	0		Negative	TN	0	0	0				No
10921373	0	0		Negative	TN	0	0	0				No

Supplementary File 3

Sample Number	FaecalMLPA GS1	FaecalMLPA GS2	FaecalTaxon-specificPCRResults	CompositeGold StandardResult	MLPA Score	CGS concisus	CGS concisusGS2	CGS jejuni	Isolate Name	Isolate MLPA Result	IsolateAdditionalResults	Discordant
10921992	0	0	Positive for GS2; negative for C. concisus	C. concisus	TP	1	0	0				Yes
10922083	0	0	Positive for C. jejuni	C. jejuni	TP	0	0	1	Cam16/54	C. jejuni	Positive for C. jejuni	No
10922840	0	0	Negative for C. concisus and GS1 & GS2	Negative	FP	0	0	0				Yes
10923701	0	0		Negative	TN	0	0	0				No
10923789	0	0	Positive for GS2; negative for Campylobacter and C. concisus	C. concisus	TP	1	0	0				Yes
10923792	0	0		Negative	TN	0	0	0				No
10924167	0	0	Positive for GS1; negative for Campylobacter and C. concisus	Negative	FP	0	0	0				Yes
10924790	0	0		Negative	TN	0	0	0				No
10925307	0	0		Negative	TN	0	0	0				No
10925990	0	0	Positive for C. concisus and GS1 & GS2; negative for Campylobacter	C. concisus	TP	1	0	0				Yes
10927618	0	0		Negative	TN	0	0	0				No
10927737	0	0		Negative	TN	0	0	0				No
10927956	0	0		Negative	TN	0	0	0				No
10928661	0	0	Negative for Campylobacter, C. concisus and GS1 & GS2	C. concisus	FN	1	0	0	Cam16/83	C. concisus GS2	Positive for C. concisus and GS2	Yes
10928732	0	0		Negative	TN	0	0	0				No
10928862	0	0		Negative	TN	0	0	0				No
10930836	0	0		Negative	TN	0	0	0				No
10930920	0	0		Negative	TN	0	0	0				No
10932176	0	0		Negative	TN	0	0	0				No
10932730	0	0		Negative	TN	0	0	0				No
10934307	0	0		Negative	TN	0	0	0				No
10934315	0	0		Negative	TN	0	0	0				No
10934320	0	0		Negative	TN	0	0	0				No
10934665	0	0		Negative	TN	0	0	0				No
10937065	0	0		Negative	TN	0	0	0				No
10937847	0	0		Negative	TN	0	0	0				No
10938313	0	0		Negative	TN	0	0	0				No
10938383	0	0		Negative	TN	0	0	0				No
10938384	0	0		Negative	TN	0	0	0				No
10939109	0	1	Positive for Campylobacter, C. concisus and GS2	C. concisus GS2	TP	1	1	0				Yes
10939423	0	0	Positive for GS1; negative for Campylobacter and C. concisus	Negative	FP	0	0	0				Yes
10939944	0	0		Negative	TN	0	0	0				No
10940091	0	0	Positive for C. jejuni	C. jejuni	TP	0	0	1	Cam16/109	C. jejuni	Positive for C. jejuni	No
10940118	0	0		Negative	TN	0	0	0				No
10940182	0	0		Negative	TN	0	0	0				No
10940507	0	1	Positive for Campylobacter and GS2; negative for C. concisus and C. coli & C. jejuni	C. jejuni + C. concisus GS2	FN	1	1	1	Cam16/118	C. jejuni	Positive for C. jejuni	Yes
10940716	0	0		Negative	TN	0	0	0				No

Supplementary File 3

Sample Number	FaecalMLPA GS1	FaecalMLPA GS2	FaecalTaxon-specificPCRResults	CompositeGold StandardResult	MLPA Score	CGS concisus	CGS concisusGS2	CGS jejuni	Isolate Name	Isolate MLPA Result	IsolateAdditionalResults	Discordant
10941593	0	1	Positive for GS1 & GS2; negative for Campylobacter and C. concisus	C. concisus GS2	TP	1	1	0				Yes
10941826	0	0		Negative	TN	0	0	0				No
10942366	0	0		Negative	TN	0	0	0				No
10942644	0	0	Positive for Campylobacter, C. concisus and GS1 & GS2	C. concisus	TP	1	0	0				Yes
10942736	0	0	Positive for C. jejuni; negative for Campylobacter	C. jejuni	FN	0	0	1	Cam16/122	C. jejuni	Positive for C. jejuni	Yes
10943202	0	0	Negative for Campylobacter, C. concisus and GS1 & GS2	Negative	FP	0	0	0				Yes
10945089	0	0		Negative	TN	0	0	0				No
10945093	0	0		Negative	TN	0	0	0				No
10945436	0	0	Positive for Campylobacter and GS2; negative for C. concisus	C. concisus	TP	1	0	0				Yes
10945895	0	0		Negative	TN	0	0	0				No
10947293	0	0		Negative	TN	0	0	0				No
10947489	0	1	Positive for Campylobacter and GS2; negative for C. concisus	C. concisus GS2	TP	1	1	0				Yes
10947883	0	0		Negative	TN	0	0	0				No
10948495	0	0		Negative	TN	0	0	0				No
10948818	0	0		Negative	TN	0	0	0				No
10949149	0	0		Negative	TN	0	0	0				No
10949912	0	0		Negative	TN	0	0	0				No
10949919	0	0		Negative	TN	0	0	0				No
10950378	0	0		Negative	TN	0	0	0				No
10951449	0	0		Negative	TN	0	0	0				No
10951760	1	1	Positive for GS1 & GS2; negative for C. concisus	C. concisus	TP	1	0	0	Cam16/152	C. concisus GS1+GS2	Positive for C. concisus and GS1	No
10952105	0	0		Negative	TN	0	0	0				No
10952514	0	0	Negative for Campylobacter, C. concisus and GS1 & GS2	C. concisus	FN	1	0	0	Cam16/159	C. concisus GS1	Positive for C. concisus and GS1	Yes
10953134	0	0		Negative	TN	0	0	0				No
10953803	0	0		Negative	TN	0	0	0				No
10953993	0	0		Negative	TN	0	0	0				No
10955153	0	1	Positive for GS1 & GS2; negative for Campylobacter and C. concisus	C. concisus GS2	TP	1	1	0				Yes
10955729	0	0	Positive for Campylobacter and C. jejuni	C. jejuni	FN	0	0	1				Yes
10957026	0	0		Negative	TN	0	0	0				No
10957317	0	0		Negative	TN	0	0	0				No
10957350	0	0		Negative	TN	0	0	0				No
10957573	0	0		Negative	TN	0	0	0				No
10957577	0	0		Negative	TN	0	0	0				No

Supplementary File 3

Sample Number	FaecalMLPA GS1	FaecalMLPA GS2	FaecalTaxon-specificPCRResults	CompositeGold StandardResult	MLPA Score	CGS concisus	CGS concisusGS2	CGS jejuni	Isolate Name	Isolate MLPA Result	IsolateAdditionalResults	Discordant
10958138	0	0		Negative	TN	0	0	0				No
10958575	0	0		Negative	TN	0	0	0				No
10958583	0	0		NA	NA	NA	NA	NA				No
10959457	NA	NA		NA	NA	NA	NA	NA				
10960053	0	0		Negative	TN	0	0	0				No
10962903	0	0		Negative	TN	0	0	0				No
10967666	0	0	Positive for C. jejuni	C. jejuni	TP	0	0	1				No
10974083	0	0		Negative	TN	0	0	0				No
10974159	0	0		Negative	TN	0	0	0				No
10974735	0	0		Negative	TN	0	0	0				No
10974997	0	0		Negative	TN	0	0	0				No
10975269	0	0		Negative	TN	0	0	0				No
10975280	0	0	Positive for GS2; negative for Campylobacter and C. concisus	C. concisus	FN	1	0	0	Cam16/301	C. concisus	Positive for C. concisus and GS1	Yes
10976591	0	1	Positive for GS2; negative for Campylobacter and C. concisus	C. concisus GS2	TP	1	1	0				Yes
10980405	0	0		Negative	TN	0	0	0				No
10981895	0	0		Negative	TN	0	0	0				No
10982518	0	1	Positive for GS2; negative for Campylobacter and C. concisus	C. concisus GS2	TP	1	1	0				Yes
10982519	0	0		Negative	TN	0	0	0				No
10983181	0	0	Positive for C. jejuni	C. jejuni	TP	0	0	1				No
10984299	0	0		Negative	TN	0	0	0				No
10984553	0	0	Positive for GS2; negative for Campylobacter and C. concisus	C. concisus	FN	1	0	0	Cam16/341	C. concisus GS2	Positive for C. concisus and GS1 & GS2	Yes
10984714	0	0		Negative	TN	0	0	0				No
10985240	0	0		Negative	TN	0	0	0				No
10985260	0	0		Negative	TN	0	0	0				No
10985295	0	0		Negative	TN	0	0	0				No
10985297	0	0		Negative	TN	0	0	0				No
10985743	0	0		Negative	TN	0	0	0				No
10985746	0	0	Positive for GS1 & GS2; negative for Campylobacter and C. concisus	C. concisus	TP	1	0	0				Yes
10986453	0	0		Negative	TN	0	0	0				No
10987715	0	0		Negative	TN	0	0	0				No
10988481	0	0		Negative	TN	0	0	0				No
10989941	0	0		Negative	TN	0	0	0				No
10996219	NA	NA		NA	NA	NA	NA	NA				
10997652	0	0		Negative	TN	0	0	0				No
10997667	0	0		Negative	TN	0	0	0				No
10997676	0	0		Negative	TN	0	0	0				No
10999803	0	0		Negative	TN	0	0	0				No
11000531	0	0		Negative	TN	0	0	0				No
11000549	0	0		Negative	TN	0	0	0				No
11000618	0	0		Negative	TN	0	0	0				No
11000841	0	0		Negative	TN	0	0	0				No

Supplementary File 3

Sample Number	FaecalMLPA GS1	FaecalMLPA GS2	FaecalTaxon-specificPCRResults	CompositeGold StandardResult	MLPA Score	CGS concisus	CGS concisusGS2	CGS jejuni	Isolate Name	Isolate MLPA Result	IsolateAdditionalResults	Discordant
11000898	0	0	Negative for A. butzleri, Campylobacter, C. coli & C. jejuni, C. concisus, GS1 & GS2, C. cuniculorum, C. helveticus, C. lari, C. upsaliensis and C. ureolyticus	A. butzleri	FN	0	0	0				Yes
11001332	0	0		Negative	TN	0	0	0				No
11003530	0	0		Negative	TN	0	0	0				No
11003706	0	0		Negative	TN	0	0	0				No
11003917	0	0		Negative	TN	0	0	0				No
11005462	0	0		Negative	TN	0	0	0				No
11005535	0	0		Negative	TN	0	0	0				No
11006998	1	1	Negative for C. concisus and GS1 & GS2	Negative	FP	0	0	0				Yes
11008004	0	0		Negative	TN	0	0	0				No
17428547	0	0	Negative for C. concisus and GS1 & GS2	Negative	FP	0	0	0				Yes
80980182	0	0		Negative	TN	0	0	0				No
80981067	0	0		Negative	TN	0	0	0				No
80981998	0	0	Negative for Campylobacter and C. coli & C. jejuni	C. jejuni	FN	0	0	1	Cam15/68	C. jejuni	Positive for C. jejuni	Yes
80982078	NA	NA		C. jejuni	NA	NA	NA	NA				
80982109	0	0		Negative	TN	0	0	0				No
80982208	NA	NA		NA	NA	NA	NA	NA				
80982211	NA	NA		NA	NA	NA	NA	NA				
80983034	NA	NA		NA	NA	NA	NA	NA				
80985548	NA	NA		NA	NA	NA	NA	NA				
80985613	NA	NA		NA	NA	NA	NA	NA				
80985689	0	0		Negative	TN	0	0	0				No
80985714	NA	NA		NA	NA	NA	NA	NA				
80985815	NA	NA		NA	NA	NA	NA	NA				
80986325	NA	NA		NA	NA	NA	NA	NA				
80987173	0	0		Negative	TN	0	0	0				No
80991380	NA	NA		NA	NA	NA	NA	NA				
80995439	0	0		Negative	TN	0	0	0				No
80998795	0	0		Negative	TN	0	0	0				No
80998837	0	0		Negative	TN	0	0	0				No
80999518	0	0		Negative	TN	0	0	0				No
81001292	0	0		Negative	TN	0	0	0				No
81001909	0	0		Negative	TN	0	0	0				No
81001910	NA	NA		NA	NA	NA	NA	NA				
81002107	NA	NA		NA	NA	NA	NA	NA				
81004068	0	0		Negative	TN	0	0	0				No
81005629	NA	NA		NA	NA	NA	NA	NA				
81005630	NA	NA		NA	NA	NA	NA	NA				
81006274	0	0		Negative	TN	0	0	0				No
81009742	0	0		Negative	TN	0	0	0				No
81012499	0	0		Negative	TN	0	0	0				No
81012500	0	0		Negative	TN	0	0	0				No

Supplementary File 3

Sample Number	FaecalMLPA GS1	FaecalMLPA GS2	FaecalTaxon-specificPCRResults	CompositeGold StandardResult	MLPA Score	CGS concisus	CGS concisusGS2	CGS jejuni	Isolate Name	Isolate MLPA Result	IsolateAdditionalResults	Discordant
81026075	0	0		Negative	TN	0	0	0				No
81026142	0	0		Negative	TN	0	0	0				No
81026219	0	0		Negative	TN	0	0	0				No
81026313	NA	NA		NA	NA	NA	NA	NA				
81026531	0	0		Negative	TN	0	0	0				No
81026553	0	0		Negative	TN	0	0	0				No
81026790	0	0		Negative	TN	0	0	0				No
81027503	0	0		Negative	TN	0	0	0				No
81027744	0	0		Negative	TN	0	0	0				No
81028359	0	0		Negative	TN	0	0	0				No
81030211	0	0	Positive for C. jejuni	C. jejuni	TP	0	0	1	Cam15/215	C. jejuni	Positive for C. jejuni	No
81033505	0	0		Negative	TN	0	0	0				No
81050641	0	0		Negative	TN	0	0	0				No
81055035	0	0		Negative	TN	0	0	0				No
81055043	NA	NA		C. concisus	NA	NA	NA	NA	Cam16/42	C. concisus GS2	Positive for C. concisus and GS1 & GS2	Yes
81055176	NA	NA		C. jejuni	NA	NA	NA	NA				
81055223	0	0		Negative	TN	0	0	0				No
81057685	0	0		Negative	TN	0	0	0				No
81058343	0	0	Positive for C. jejuni	C. jejuni	TP	0	0	1	Cam16/49	C. jejuni	Positive for C. jejuni	No
81058472	0	0	Positive for C. jejuni	C. jejuni	TP	0	0	1	Cam16/48	C. jejuni	Positive for C. jejuni	No
81059654	0	0	Positive for C. jejuni; negative for Campylobacter	C. jejuni	FN	0	0	1	Cam16/56	C. jejuni	Positive for C. jejuni	Yes
81060271	0	0	Positive for C. jejuni	C. jejuni	TP	0	0	1	Cam16/53	Not undertaken		No
81061452	0	0		Negative	TN	0	0	0				No
81061562	0	0	Positive for C. coli & C. jejuni	C. coli	TP	0	0	0	Cam16/55	C. coli	Positive for C. coli	No
81061807	0	0		Negative	TN	0	0	0				No
81065359	0	0		Negative	TN	0	0	0				No
81065939	0	0		Negative	TN	0	0	0				No
81067085	0	0		Negative	TN	0	0	0				No
81067880	0	0		Negative	TN	0	0	0				No
81074287	0	0		Negative	TN	0	0	0				No
81075799	NA	NA		NA	NA	NA	NA	NA				
81076392	0	0		Negative	TN	0	0	0				No
81076609	0	0		Negative	TN	0	0	0				No
81078259	0	0		Negative	TN	0	0	0				No

Supplementary File 3

Sample Number	FaecalMLPA GS1	FaecalMLPA GS2	FaecalTaxon-specificPCRResults	CompositeGold StandardResult	MLPA Score	CGS concisus	CGS concisusGS2	CGS jejuni	Isolate Name	Isolate MLPA Result	IsolateAdditionalResults	Discordant
81078872	0	0	Positive for GS1 and C. jejuni; negative for Campylobacter and C. concisus	C. jejuni	TP	0	0	1	Cam16/131	C. jejuni	Positive for C. jejuni	Yes
81079425	0	0		Negative	TN	0	0	0				No
81081451	0	1	Positive for Campylobacter and GS1 & GS2; negative for C. concisus	C. concisus GS2	TP	1	1	0				Yes
81081631	0	0		Negative	TN	0	0	0				No
81082674	0	0		Negative	TN	0	0	0				No
81085323	0	0		Negative	TN	0	0	0				No
81087854	0	0		Negative	TN	0	0	0				No
81088350	0	0	Positive for Campylobacter and C. jejuni	C. jejuni	TP	0	0	1				Yes
81088435	0	0		Negative	TN	0	0	0				No
81088706	0	0		Negative	TN	0	0	0				No
81090488	0	0		Negative	TN	0	0	0				No
81091372	0	0		Negative	TN	0	0	0				No
81094123	0	0	Positive for C. jejuni	C. jejuni	TP	0	0	1				No
81094482	0	0		Negative	TN	0	0	0				No
81098106	0	0		Negative	TN	0	0	0				No
81100344	0	0	Positive for C. jejuni	C. jejuni	TP	0	0	1				No
81101045	0	0	Positive for Campylobacter and GS1 & GS2; negative for C. concisus	C. concisus	FN	1	0	0	Cam16/303	C. concisus GS1	Positive for C. concisus and GS1	Yes
81102268	0	0		Negative	TN	0	0	0				No
81102337	0	0		Negative	TN	0	0	0				No
81102587	0	0		Negative	TN	0	0	0				No
81102588	0	0		Negative	TN	0	0	0				No
81103417	0	0		Negative	TN	0	0	0				No
81104380	0	0		Negative	TN	0	0	0				No
81105862	NA	NA		NA	NA	NA	NA	NA				
81106009	0	0		Negative	TN	0	0	0				No
81107814	0	0		Negative	TN	0	0	0				No
81110257	0	0		Negative	TN	0	0	0				No
81112309	0	0		Negative	TN	0	0	0				No
81113260	0	1	Positive for Campylobacter, C. concisus and GS1 & GS2	C. concisus GS2	TP	1	1	0				Yes
81113568	0	0		Negative	TN	0	0	0				No
81114620	0	1	Positive for Campylobacter, C. concisus and GS2	C. concisus GS2	TP	1	1	0				Yes
81115274	0	0		Negative	TN	0	0	0				No
81115605	0	0		Negative	TN	0	0	0				No